PREDICTION AND ANALYSIS OF THE METHYLATION STATUS OF CPG ISLANDS IN HUMAN GENOME

A Thesis
Presented to
The Academic Faculty

by

Hao Zheng

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Electrical and Computer Engineering

Georgia Institute of Technology
May 2012

PREDICTION AND ANALYSIS OF THE METHYLATION STATUS OF CPG ISLANDS IN HUMAN GENOME

Approved by:

Hongwei Wu, Advisor
School of Electrical and Computer Engineering
Georgia Institute of Technology

Chris Barnes
School of Electrical and Computer Engineering
Georgia Institute of Technology

Elliot Moore II
School of Electrical and Computer Engineering
Georgia Institute of Technology

King Jordan
School of Biology
Georgia Institute of Technology

Shi-Wen (Albert) Jiang
Department of Biomedical Sciences
Mercer University School of Medicine

Date Approved: March 16, 2012

This thesis is dedicated to my parents for their unconditional love and endless support.

ACKNOWLEDGEMENTS

This dissertation would not have been possible without the help of numerous people. I owe my gratitude to all these people.

First, I would like to express my gratitude to Dr. Hongwei Wu, my advisor. You have continually guided and encouraged me to conduct more research since I first entered the Ph.D. program. You are instrumental in my development as a bioinformatics researcher. Four years ago, I would have never guessed I would be where I am today. You are an awesome advisor, and for this, I thank you.

I also would like to thank my thesis committee members, Dr. Chris Barnes, Dr. Elliot Moore, Dr. King Jordan, and Dr. Shi-Wen Jiang for their precious time and valuable suggestions for the work done in this dissertation.

Next, I would also like to thank my fellow lab members and my friends who provided great collaboration and assistance during my study. You have also made my long journey much more cheerful.

Finally, my special appreciations go out to my parents to whom I owe so much. My parents were very supportive of me during my research and always encouraged me to aim higher in everything I do.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

DNA methylation serves as a major epigenetic modification crucial to the normal organismal development and the onset and progression of complex diseases such as cancer. Computational predictions for DNA methylation profiling serve multiple purposes. First, accurate predictions can contribute valuable information for speeding up genome-wide DNA methylation profiling so that experimental resources can be focused on a few selected while computational procedures are applied to the bulk of the genome. Second, computational predictions can extract functional features and construct useful models of DNA methylation based on existing data, and can therefore be used as an initial step toward quantitative identification of critical factors or pathways controlling DNA methylation patterns. Third, computational prediction of DNA methylation can provide benchmark data to calibrate DNA methylation profiling equipment and to consolidate profiling results from different equipments or techniques.

This thesis is written based on our study on the computational analysis of the DNA methylation patterns of the human genome. Particularly, we have established computational models (1) to predict the methylation patterns of the CpG islands in normal conditions, and (2) to detect the CpG islands that are unmethylated in normal conditions but aberrantly methylated in cancer conditions. When evaluated using the CD4 lymphocyte data of Human Epigenome Project (HEP) data set based on bisulfite sequencing, our computational models for predicting the methylation status of CpG islands in the normal conditions can achieve a high accuracy of 93-94%, specificity of 94%, and sensitivity of 92-93%. And, when evaluated using the aberrant methylation data from the MethCancerDB database for aberrantly methylated genes in cancer, our models for detecting the CpG islands that are unmethylated in normal conditions but aberrantly methylated in colon or prostate cancer can achieve an accuracy of 92-93%, specificity of 98-99%, and sensitivity of 92-93%.

The contribution of this thesis lies in three aspects. First, we identify various genetic and epigenetic features that are associated with the methylation status and cancer related aberrant methylation of the CpG islands in human chromosomes. These features provide the foundation for exploring the exact mechanisms of DNA methylation in normal organismal development and cancerogenesis. Second, our DNA methylation predictive model serves as a fast and effective way to explore genome-wide CpG island methylation profiles in normal tissues. Third, our predictive model for cancer related aberrant methylation can be used as an initial step to prioritize and detect novel aberrantly methylated genes in cancer.

CHAPTER I

INTRODUCTION

Epigenetics refers to a somatically inheritable pattern of gene expression that is determined by mechanisms other than those encoded in DNA sequences. DNA methylation is an important type of epigenetic modification, implicated in critical cellular functions including genetic imprinting, X-chromosome inactivation, suppression of retroviral elements, and carcinogenesis. In mammals, DNA methylation involves the addition of a methyl group to DNA via DNA methyltransferase (DNMT), and typically occurs at the cytosine residues in a CpG dinucleotide context [1][2].

CpG dinucleotides in human genome are relatively rare but are enriched in short DNA segments known as CpG islands (CGIs) [3]. Most CpG dinucleotides are methylated in human somatic cells [4], but the CpG dinucleotides residing within CGIs tend to remain unmethylated. A CGI is traditionally defined as a stretch of DNA sequence that fulfills the Gardiner-Garden criteria: (*i*) with $\geq$200 base pairs (bps), (*ii*) with a GC content >50%, and (*iii*) with an observed/expected CpG ratio $\geq$60% [5]. CGIs often colocalize with functional promoter regions, and the methylation status of CGI serves as an important mechanism for epigenetic gene control. In human genome, CGIs overlap with the promoter regions of approximately 50-60% of known genes, including most housekeeping genes [6].

On the one hand, DNA methylation can be determined experimentally using biochemical experiment-based approaches. On the other hand, computational modeling can effectively complement the wet chemistry approach to identify critical features or pathways controlling DNA methylation patterns, to provide valuable information when the DNA methylation data are unavailable for certain genomic regions, as well as to calibrate DNA methylation profiling equipment and to consolidate profiling results from different equipments or techniques.

We perform computational analysis of the DNA methylation and aberrant methylation

patterns in human genome. The objective of this work is to (a) identify various features that are associated with the methylation status of CGIs in normal tissues and aberrant methylation of CGIs in cancerous conditions, (b) discriminate between CGIs that are prone to methylation from those that are resistant to methylation based on the identified features, and (c) distinguish the aberrantly methylated CGIs in cancer from those that are consistently unmethylated.

The outline of the remaining document is as follows.

Chapter II introduces the biological background of DNA methylation, followed by a brief review of biochemical experiment-based DNA methylation profiling techniques, and summary of major existing DNA methylation data sets and databases. Then, computational modeling for DNA methylation and our contributions are introduced.

Chapter III presents a schematic overview of the workflow designed for predicting methylation status of CGIs in normal tissues, and detecting cancer related aberrant methylation in our research.

Chapter IV introduces how we incorporate various resources to form the training data for CGI methylation prediction in normal tissues, and the training data for cancer related aberrant methylation prediction.

Chapter V presents details of our method developed and implemented for predicting CGI methylation and cancer related aberrant methylation. The model development consists of three core parts – feature extraction, feature selection and model construction through prediction tests.

Chapter VI provides the results and discussions of our experiments, including (*i*) various genetic and epigenetic features that have been identified to be associated with the methylation status of the CGIs in normal tissues and CGI aberrant methylation in cancer, (*ii*) performance of our predictive models for CGI methylation in normal tissues, and (*iii*) performance of our predictive model to detect potential CGIs that are aberrantly methylated in cancer.

Chapter VII draws conclusions and discusses possible future directions for DNA methylation.

Chapter VIII includes the appendices for the supplementary materials, including the abbreviated terms used in this thesis and the results of enrichment analysis.

CHAPTER II

ORIGIN AND HISTORY OF THE PROBLEM

2.1 Biological Background

Epigenetics is the study of heritable changes in genotypes or phenotypes caused by mechanisms that are not encoded in the underlying DNA sequence [7]. DNA methylation is a type of epigenetic modification which involves the addition of a methyl group to DNA via DNMT. In mammals, DNA methylation typically occurs at the cytosine residues in a CpG dinucleotide context (i.e., a cytosine directly followed by a guanine) [1]. The "p" in a CpG dinucleotide denotes the phosphodiester bond between the cytosine and the guanine residues. Generally, CpG dinucleotides are observed to be relatively rare in most sequenced mammalian genomes (observed/expected CpG ratio: ∼20%-25%), which is mainly attributed to the hypermutability of methylated CpG to TpG or CpA in the complementary strand [8]. However, CpG dinucleotides are enriched in short DNA segments known as CpG islands (CGIs), as compared to bulk DNA [9].

Traditionally, a CGI is defined as a sequence region that fulfills the Gardiner-Garden criteria: (i) at least 200 base pairs (bps), (ii) with a GC content that is greater than 50%, and (iii) with an observed/expected CpG ratio that is greater than 60% [5]. CGIs are generally located around the 5' end of genes and considered as gene markers [10]. Most (70-80%) of the CpG dinucleotides are generally methylated in human somatic cells [4]. However, unlike the global methylation of CpGs in the bulk of the genome, the CpG dinucleotides residing within CGIs tend to be unmethylated.

DNA methylation plays an important role for gene control during normal cell development and cell differentiation. For example, it has been demonstrated that DNA methylation, together with DNMT1 (one type of DNMT), is instrumental to the regulation of gene expression in T cells during the cell development stage [11]. And, in differentiated cells, genes encoding pluripotency transcription factors are suppressed by DNA methylation so

as to avoid de-differentiation [12][13]. DNA methylation is also associated with a number of key processes including genomic imprinting [14], X-chromosome inactivation in females [15], maintenance of repetitive elements [16], and tumorigenesis [17]. For example, hypermethylation of tumor suppressor genes (TSGs) and hypomethylation of oncogenes are associated with various cancers, including sporadic retinoblastoma [18], prostate cancer [19], liver cancer [20] and salivary gland adenoid cystic carcinoma [21]. Numerous studies have suggested that DNA methylation patterns can be used for early detection and precise sub-typing of cancers, and to predict and monitor drug/therapeutic effects [22, 6, 23].

As a result, it is helpful to study the different methylation patterns of CGIs around certain key genomic regions (e.g. cancer-related genes) compared with other genomic regions (e.g. constitutively unmethylated regions). However, all these studies of differential methylation are very much impeded by the lack of genome-wide and tissue-specific CGI methylation data. Due to the above reasons, the profiling and prediction of DNA methylation status of CGIs in human genomes has become one of the most pressing and important topics in computational biology recently.

2.2 Biochemical Experiment-based DNA Methylation Profiling

The biochemical experiment-based approaches for characterizing DNA methylation are mainly based on bisulfite conversion, methylation-specific restriction and immunoprecipitation, respectively [24]. These approaches are usually accompanied by array-based or high-throughput sequencing-based DNA methylation analysis. We here briefly describe these approaches [25].

Bisulfite conversion: When the DNA is treated with sodium bisulfite, the unmethylated cytosines are converted to uracils and then to thymines during polymerase chain reaction (PCR) amplifications, while the methylated cytosines remain unchanged [26]. The bisulfite treated DNA can then be interrogated by array hybridization (e.g., high density tiling array-based genome-wide DNA methylation profiling [27]) or DNA sequencing to determine which CpG dinucleotides are methylated. Particularly, with the advance of sequencing techniques (e.g., Illumina sequencing [28][29], Roche 454 pyrosequencing [30] and

ABI SOLiD sequencing [31]), bisulfite sequencing has gained its popularity and is becoming the gold standard in detecting DNA methylation [32][33]. Basically, regions of interest are amplified and cloned, then clones are sequenced to confirm methylation (presence of cytosine) or no methylation (presence of thymine at a known CpG). Because the bisulfite treatment introduces changes to the DNA sequence that depend on the methylation status of individual cytosine residues, this approach can potentially yield single-nucleotide resolution information about the methylation status of the DNA.

Methylation specific restriction enzyme digestion: The biochemical foundation of this approach is that some restriction enzymes (methylation dependent restriction enzymes, MDREs) are able to, while some other restriction enzymes (methylation sensitive restriction enzymes, MSREs) are unable to cleave methylated cytosines in their recognition sites. For instance, HpaII, a commonly used MSRE, recognizes CCGG and is unable to cut DNA when the internal cytosine is methylated [34]. McrBC, a commonly used MDRE, recognizes two half sites of the form $(G/A)^mC$ and does not act upon unmethylated DNA [35]. The Restriction Enzyme Database (REBASE) contains a more complete list and detailed information of MDREs and MSREs [36]. After DNA digestion using MDREs or MSREs, further processing is needed to derive the DNA methylation information. The first way of deriving the DNA methylation information is based on the distribution of the length of the digested fragments – methylated DNA that cannot be digested tend to result in longer fragments [37]. The second way is based on real time quantitative PCR (qPCR) with specially designed primers. For example, by using the primers that are specifically designed to target hypermethylated DNA fragments, Ng, *et al.*, were able to calculate the difference between the cycle threshold values of the methylated and unmethylated fragments and then estimate the percentage of methylation [38]. Real-time qPCR has gained its popularity for quantifying the methylation status recently [39][40]. Most recently, Ng, *et al.*, developed a MSRE-based quantitative assay to measure percentage of methylation in plasma [38]. No matter digested or not by restriction enzyme, the remaining DNA was quantified by qPCR using primers flanking the hypermethylated promoter CpG region. Percentage of methylation in tissue samples was calculated by the difference of the cycle threshold

(Ct) values from methylated signal and Ct values from unmethylated signal formulated as: $100/[1 + 2^{\Delta Ct_{(meth-unmeth)}}]\%$ [38]. Restriction-based methods, together with microarrays, can also be used to provide localization information of DNA methylation [41]. The third way is based on micro-array or sequencing [41]. For example, Edwards, *et al.*, developed a method called Methyl-MAPS (methylation mapping analysis by paired-end sequencing), which combines MSREs (AciI, HhaI, HpaII, HpyCH4IV, and BstUI) and MDRE (McrBC)-based restriction with deep sequencing, and could achieve single-CpG resolution and cover up to 82.4% of all the CpG dinucleotides in the human and mouse genomes [42].

Methylated DNA immunoprecipitation (MeDIP): MeDIP is an efficient large-scale technique that uses antibodies raised against methylated cytosines to enrich methylated DNA fragments [43]. It consists of four steps. First, DNA is extracted, purified, randomly sheared by sonication into fragments of 300-1,000 bps, denatured, and immunoprecipitated with an antibody that detects 5-methylcytidine (5mC antibody) [44]. Secondly, endopeptidase K is used to digest the antibodies and leave the methylated DNA intact. Thirdly, phenol/chloroform is used to remove the digested proteins, and precipitation purifies the DNA [45]. Finally, the resulting purified methylated DNA fragments can then be used for methylation studies by locus-specific PCRs [45], microarrays (MeDIP-chip) [46][47] or sequencing (MeDIP-seq) [48].

Among these three techniques, bisulfite sequencing is the only one that provides single-nucleotide resolution and covers the vast majority (>90%) of the genome. However, this technique is more expensive than the other two, and suffers the drawback that the conversion of unmethylated cytosines to uracils can be unstable [49]. Methylation specific restriction enzyme digestion is generally simpler and faster to establish and requires no base modification. The disadvantage of this technique lies in that not all CpGs are located within the recognition sites of the restriction enzymes, and that it gives rise to false-positives if the enzyme digestion is not complete [50]. However, the recently emerged Methyl-MAPS, which combines MSRE- and MDRE-based restriction with deep sequencing, could achieve single CpG resolution and cover $\leq$ 82.4% of all the CpG dinucletoides in the human and

mouse genomes. Although its coverage is a bit worse than the bisulfite sequencing techniques, Methyl-MAPS is substantially cheaper and applicable to more repetitive regions, and is therefore a very viable solution to the high resolution genome-wide DNA methylation profiling [42]. The MeDIP methods have relatively lower resolution (a few hundred base pairs at best), but less sequence bias than the other two techniques [24].

2.3 DNA Methylation Data Sets and Databases

2.3.1 Data sets

The Human Epigenome Project (HEP) was officially launched in 2003 by the Wellcome Trust Sanger Institute, Epigenomics AG, and the Center National de Génotypage in Europe [51]. HEP aims to identify, catalogue and interpret genome-wide DNA methylation patterns of all human genes in all major tissues [52]. Within this context, large scaled data sets have been made available to archive the DNA methylation profiles in various tissues or cells. DNA methylation profiling of the Human Major Histocompatibility Complex, the most gene-dense region in the human genome containing genes with a diversity of functions on chromosome 6 (6p21.3), was one of the first studies in HEP [53]. Eckhardt *et al.* later created by far the most comprehensive and updated HEP data set by profiling 1.9 million CpG dinucleotides of chromosomes 6, 20 and 22 of 43 samples from 12 different normal human tissues [54].

Beyond HEP there are several other large-scaled data sets of DNA methylation, including Lister, *et al.*'s [29] and Bell, *et al.*'s [55] that are based on bisulfite treatment, methylation profiles of DNA (mPod) [48][47] and Weber, *et al.*'s [43][46] that are based on MeDIP, as well as Methyl-MAPS [42], Kaminsky, *et al.*'s [56], Schumacher, *et al.*'s [57], and Flanagan, *et al.*'s [58] that are based on methylation specific restriction enzyme digestion. Lister, *et al.*'s data set is for human embryonic stem cell and fetal fibroblasts, and provides the first genome-wide maps of methylated cytosines at the resolution of single nucleotides. It covers ∼62 million and ∼45 million methylcytosines of all chromosomes in the embryonic stem cell and fetal fibroblasts, respectively [29]. Bell, *et al.*'s data set is for lymphoblastoid cell lines from 77 HapMap Yoruba individuals, and contains the methylation measurements

of 22,290 CpG dinucleotides of all chromosomes [55]. mPod is based on a combination of MeDIP, microarray and bioinformatic analysis, and contains the methylation status of 69,510 genomic regions (each ∼500 bps long) of 13 normal somatic tissues, placenta, sperm and the GM06990 immortalized cell line [48][47]. Weber, *et al.*, created two data sets, one covering ∼6,000 CGIs of the primary fibroblasts, and the other covering ∼16,000 promoters in primary somatic and germline cells [43][46]. Methyl-MAPS data set contains the methylation status of 152,693,954 CpG dinucleotides for breast, and 75,676,854 CpG dinucleotides in for brain [42]. Kaminsky, *et al.*'s data set contains the methylation status of 12,323 loci in the white blood cells and buccal epithelial cells of monozygotic and dizygotic twins [56]. Schumacher, *et al.*'s data set contains the methylation status of CpG dinucleotides in chromosomes 21 and 22 for prefontal cortex tissues of eight individuals, covering $\sim 0.1\%$ of the human genome [57]. Flanagan, *et al.*'s methylation data set covers ∼4,970 unique loci of all chromosomes of the germ cell [58]. Schumacher *et al.*'s and Flanagan *et al.*'s data sets are collected by MethyLogiX [59].

We summarize in Table 1 the cell, tissue or phenotypes of these data sets, as well as their coverage, resolution, and underlying biochemical approaches. For other relevant but non-human DNA methylation data sets (such as those for the mouse genome), the readers are referred to [60, 28, 61] and the references therein.

2.3.2 Databases

In addition to the data sets, databases are being constructed to archive the DNA methylation profiles and to link such information with other genotypic and phenotypic information. Such databases include MethPrimerDB [62], MethDB [63], MethCancerDB [64], PubMeth[65], MeInfoText[66], and MethyCancer [67].

MethPrimerDB (`http://medgen.ugent.be/methprimerdb/`) is a database of PCR primers for DNA methylation profiling experiments [62]. So far, 259 primer sets contributed by 135 different resources that have been validated using PCR-based methylation assays are available.

MethDB (`http://www.methdb.de/`) aggregates and attempts to standardize the DNA

Table 1: Summary of available major human genomic DNA methylation data sets and their information

Data Set	Cell, Tissue, Phenotype and Sample	Coverage and Resolution	Biochemical Method	URL
HEP [54]	12 normal tissues: CD4+ and CD8+ lymphocytes, dermal fibroblasts, dermal keratinocytes, dermal melanocytes, fetal liver, fetal skeletal muscle, heart muscle, liver, placenta, skeletal muscle, sperm; 43 samples	1.9 million CpG dinucleotide from chromosomes 6, 20 and 22	bisulfite sequencing	http://www.ucl.ac.uk/cancer/research-groups/medical-genomics/hep_data/
Lister *et al's*[29]	human embryonic stem cells and fetal fibroblasts	genome-wide, single-base-resolution	bisulfite sequencing	http://neomorph.salk.edu/human_methylome
Bell *et al's* [55]	lymphoblastoid cell lines from 77 HapMap Yoruba individuals	22,290 CpG dinucleotides across all chromosomes	bisulfite conversion based	http://eqtl.uchicago.edu/Methylation/
mPod [47][48]	normal B-cells, CD8 T-cells, CD4 T-cells, cervix, colon, liver, lung, rectum, pancreas, prostate, placenta, skeletal muscle, sperm, uterus, whole blood, and the EBV-transformed GM06690 cell line.	69,510 genomic regions of 500 bps long; genome-wide	MeDIP-chip	ftp://ftp.ebi.ac.uk/pub/software/ensembl/efg/MeDIP-chip/
Weber05 [43]	primary fibroblasts, colon cancer cells, normal colon mucosa	∼6,000 CGIs across all chromosomes	MeDIP	http://www.fmi.ch/groups/schubeler.d/web/data.html
Weber07 [46]	primary fibroblasts and sperm cells	16,000 promoters across all chromosomes	MeDIP-chip	http://www.fmi.ch/groups/schubeler.d/web/data.html
Methyl-MAPS [42]	normal breast and brain	up to 82.4% of the CpG dinucleotides in the genome	methylation sensitive and dependent restriction enzymes	http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi
Kaminsky *et al's* [56]	white blood cells, buccal epithelial cells and gut biopsies of 114 monozygotic twins; white blood cells and buccal epithelial cells of 80 dizygotic twins	6,405 white blood cell loci, 5,918 buccal cell loci, and 5,941 gut biopsies loci of 22 chromosomes	methylation sensitive restriction enzyme	http://camh.net/Research/Areas_of_research/Epigenomics/twin_data.html
Schumacher *et al's* [57]	prefontal cortex tissues in eight individuals	$\sim 0.1\%$ of the human genome	methylation sensitive restriction enzyme	http://www.methylogix.com/genetics/database.shtml.htm
Flanagan *et al's* [58]	germ cells from 46 individuals	∼4,970 unique loci across all chromosomes	methylation sensitive restriction enzyme	http://www.methylogix.com/genetics/database.shtml.htm

methylation data from multiple resources. It currently contains over 19,905 methylation content data and 5,382 methylation patterns or profiles for 48 species, 1,511 individuals, 198 tissues and cell lines, and 79 phenotypes [63]. Here, methylation content refers to the percentage of methylated cytosines in the genome (without position information) [68]; methylation pattern refers to the methylation status of a series of CpG dinucleotides that are located in close proximity and form a CpG-rich region [69]; methylation profile refers to the methylation status of all cytosines in the genome [69]; and the phenotypes include healthy, tumor, rheumatoid arthritis, etc.

MethCancerDB, PubMeth, MeInfoText, and MethyCancer are four major databases containing cancer-related methylation information. MethCancerDB (`http://www.methcancerdb.net/methcancerdb/home.seam`) contains the data collected from over 300 resources about cancer-related aberrant CpG methylation. It focuses on the CGIs around genes (currently covering 2,199 genes) and experimental designs such as diagnosis and prognosis [64]. PubMeth (`http://mit.lifescience.ntu.edu.tw/`) is based on literature search, and contains over 440 genes that are reported to be methylated in over 43 cancer types [65]. It concentrates on methylation frequency of genes in cancer samples without systematically distinguishing between cancer subphenotypes [70]. MeInfoText presents the profile of gene methylation among over 205 human cancer types based on association mining from large amounts of literature [66]. MethyCancer (`http://methycancer.genomics.org.cn`) integrates data from public resources (e.g., MethDB and HEP) and from data produced from China's Cancer Epigenome Project. It currently contains over 485 annotated cancer genes with methylation data from 511 cancer types [67]. We summarize in Table 2 the database name, number of tissue/primer/cancer types, coverage, as well as their sources and references.

2.4 Computational Modeling for DNA Methylation

A classical view is that CpG dinucleotides inside promoter-related CGIs are generally unmethylated in normal tissues [71]. However, some pilot studies on methylation status of CGIs show that a sizable fraction of CGIs are actually methylated in normal tissues [72][43].

Table 2: Summary of existing major human genomic DNA methylation databases and their main contribution

Database Name	No. Tissue /Primer/Cancer Type	Coverage	Source	Reference	url
MethDB	198 tissues and cell lines	19,905 methylation content data and 5,382 methylation patterns or profiles for 48 species	literature; submission	[63]	http://www.methdb.de/
methPrimerDB	over 259 PCR primers for popular DNA methylation analysis methods	NA	submitted by 135 people	[62]	http://medgen.ugent.be/methprimerdb/
MethCancerDB	48 cancer types	over 2,199 genes	348 sources	[64]	http://www.methcancerdb.net/methcancerdb/home.seam
PubMeth	43 cancer types	over 440 genes	over 1,000 literatures	[65]	http://www.pubmeth.org/
MeInfoText	over 205 cancer types	NA	literature mining	[66]	http://mit.lifescience.ntu.edu.tw/
MethyCancer	511 cancer types	485 genes	MethDB, HEP, Cancer Epigenome Project in China, etc.	[67]	http://methycancer.genomics.org.cn

It is still unclear what mechanisms determine certain CGIs to be methylated while others not, and it has become more and more interesting to profile the methylation status of CGIs in human genomes. In addition to biochemical experiment-based techniques for DNA methylation profiling, computational prediction of DNA methylation has been carried out by numerous researchers. Such computational predictions serve multiple purposes. First, accurate predictions can contribute valuable information for speeding up genome-wide DNA methylation profiling so that experimental resources can be focused on a selected few, while computational procedures are applied to the bulk of the genome [6]. Second, computational predictions can extract functional features and construct useful models of DNA methylation based on existing data. These can be used as an initial step toward quantitative identification of critical factors or pathways controlling DNA methylation patterns [73]. Third, computational prediction of DNA methylation can provide benchmark data to calibrate DNA methylation profiling equipment and to consolidate profiling results from different equipments or techniques.

2.4.1 Statistical Measurements

The performance of a computational predictive model is commonly assessed by using three statistical measurements defined in Eqns. (1)–(3), namely, sensitivity (SE), specificity (SP), and accuracy (ACC),

$$SE = \frac{TP}{TP + FN}, \tag{1}$$

$$SP = \frac{TN}{TN + FP}, \tag{2}$$

$$ACC = \frac{TP + TN}{TP + FN + TN + FP}, \tag{3}$$

where TP, TN, FP and FN represent true-positives, true-negatives, false-positives, and false-negatives, respectively. The definition of the positive and negative data are problem-dependent [74].

2.4.2 Existing Models

Computational predictive models have been developed to identify CpG dinucleotides methylated or unmethylated [75][76], CGIs (or CpG-rich segments) methylated or unmethylated [74][6][77][78][79], and CGIs (or CpG-rich regions) that are differentially methylated in different tissues/cell types or phenotypes [9][80][81].

A key step for building computational predictive models is to select informative features. For the prediction of DNA methylation, the features can be roughly grouped into two broad categories: genetic and epigenetic. Given a region of interest (ROI), e.g., a CGI or a genomic region centered around a particular CpG dinucleotide, the genetic features include:

1. general attributes of the ROI (e.g., length of the ROI, and distribution of the CpG dinucleotides in the ROI);
2. patterns of the DNA composition of the ROI;
3. patterns of functional or conserved elements within or near the ROI;
4. structural and physicochemical properties of the ROI;
5. functions of the genes within or near the ROI;

6. the extent of the diversity of the ROI within the population;

7. the extent of the conservation of the ROI among species;

The epigenetic features mainly regard the methylation and acetylation status of histones, proteins serving as spools around which DNA winds and playing important roles in gene regulation.

Bhasin, *et al.*, used DNA composition features to predict the methylation of single cytosines. A 39-nucleotide long DNA fragment centered around the cytosine of interest was considered as the ROI, and each nucleotide in the ROI was coded by using a 5-bit binary sparse code. In this way, each ROI was represented by a series of codes, and the difference between ROIs was able to be quantified. An accuracy of about 75% was reported using a support vector machine-based classifier [75]. Lu, *et al.*, also used DNA composition features for predicting whether a CpG dinucleotide is methylated or not. A 1,000 nucleotide long DNA fragment centered around the CpG dinucleotide was used as the ROI, and the frequencies of all pentamer oligonucleotides formed the features. An accuracy of about 77% was reported for the CD4 lymphocytes data set using a nearest neighbor-based classifier [76]. Feltus, *et al.*, used frequencies of seven DNA patterns, TCCCCCNC, TTTCCTNC, TCCNCCNCCC, GGAGNAAG, GAGANAAG, GCCACCCC, and GAGGAGGNNG, with N representing any base, and achieved an accuracy of about 82% on the human fibroblast data set when distinguishing between methylation-prone and methylation-resistant CGIs using a linear programming-based classifier [9].

In addition to DNA composition features, Fang, *et al.*, also used the distribution of the repetitive element AluY, as well as the distribution of TFBSs, for predicting the methylation status of CpG rich segments. They reported a specificity of about 84% and sensitivity of about 84% on the human brain data set using a support vector machine-based classifier [74]. Bock, *et al.*, used DNA composition features, predicted DNA helix structure, attributes of repeat elements and TFBSs, evolutionary conservation of PhastCons elements [82] and the number of single nucleotide polymorphisms (SNPs) for the prediction of CGI methylation

[6][77], and their method achieved a high specificity (about 98%) but a relatively low sensitivity (about 67%) on human lymphocytes using a support vector machine-based classifier [79]. Ali, *et al.*, also used the DNA composition information, predicted DNA structure, and SNP features, and reported an accuracy of about 72% on the human lymphocytes data set using a K nearest neighbor-based classifier [78]. To predict tissue-specific differentially methylated regions (DMRs), Previti, *et al.*, used CGI specific attributes, attributes of repetitive elements, number and frequency of PhastCons elements, as well as structural and physicochemical properties. When classifying CGIs into four categories: constitutively methylated, constitutively unmethylated, tissue-specific DMR, and lack of methylation exclusively in sperm, they reported an accuracy of about 89% using a decision tree-based classifier [80]. Lv, *et al.*, detected novel hypermethylated genes in breast cancer with area under the receiver operating characteristic (ROC) curve (sensitivity vs. 1-specificity) larger than 0.7 [81].

Computational prediction models that are solely based on genetic features can partially characterize DNA methylation status. This is because DNA methylation, as an epigenetic phenomenon, is affected by some other epigenetic factors, such as histone methylation and histone acetylation. In light of the reported interaction between histone modification enzymes and DNA methylases [83][46], Fan, *et al.*, found four histone methylation marks that are highly correlated with the DNA methylation status of CGIs, and then incorporated these histone methylation marks into the prediction of the methylation status of CGIs. Compared to those methods without histone methylation information [79][77], the augmented features indeed led to improved performance: a specificity of about 94% and a sensitivity of about 74% on the CD4 T cell data set using a support vector machine-based classifier [79].

2.4.3 Contribution

In this study, we analyze cancer-related aberrant DNA methylation, to identify patterns indicative of methylation variation in normal versus cancerous cells. We identify those CGIs that are methylated in human tumors, but are unmethylated in normal tissues. A crucial step is to construct a high-resolution CGI methylation map genome-wide. Thus,

we first investigate the association between the CGI methylation and various potentially methylation-related feature classes. Our feature classes can be divided into eight categories:

1. CGI specific attributes;
2. DNA sequence patterns;
3. DNA structure patterns;
4. distribution of TFBSs;
5. distribution of the evolutionarily conserved elements;
6. gene functions;
7. histone methylation status;
8. histone acetylation status;

We select a subset of these features that show the most predictive power between methylated and unmethylated CGIs by virtue of sequence analysis techniques and statistical tests, and further build a binary classifier as well as a regression model with machine learning techniques to construct the genome-wide map of the methylation status of CGIs. Finally, we constructed novel models to detect those CGIs that are potentially to be subject to aberrant methylation in different cancer types.

We identify 342 features that are statistically significantly associated with CGI methylation in CD4 T cell. These features span across all the eight feature categories. We use principal component analysis (PCA) to further decorrelate these features and build models for predicting binary CGI methylation status in normal CD4 T cell. Our models can achieve an accuracy of about 93-94% , specificity of about 94% , and sensitivity of about 92-93%. We also demonstrate that our models can have high generalizability to other normal tissues and cell types as well. We also design regression models to profile the CGI methylation in normal CD4 T cell instead of binary prediction. For cancer related aberrant DNA methylation, we identify 88 features that are correlated with CGI differential methylation in cancer. We also detect 75 and 45 discriminant features for aberrant methylation in

colon and prostate cancer, respectively. Furthermore, based on these signature features, we construct models that can achieve high accuracy (∼92%-∼93%), specificity (∼98%-∼99%) as well as sensitivity (∼92%-∼93%) for predicting aberrantly methylated genes in both the colon and prostate cancer, using housekeeping genes as a negative control group. We also apply our computational models to all promoter CGIs in human genome to infer and prioritize novel aberrantly methylated genes in cancer.

CHAPTER III

OVERVIEW OF THE WORKFLOW

We aim to (1) develop computational predictive models to profile CGI methylation in normal tissues and (2) to identify those CGIs that have high potential to be aberrantly methylated in human cancers, but are unmethylated in normal tissues. We design the workflow for constructing such predictive models, as illustrated in Fig. 1. We first develop a computational model using the HEP data set (which is based on the bisulfite sequencing technology and is therefore of high resolution but low coverage) to provide the methylation status of all the CGIs of the human genome. We then screen out those CGIs that are methylated in cancerous conditions (as indicated by MethCancerDB) but unmethylated in normal conditions (according to the mPod data set and our computational predictions) to form the positive data set; and screen out those CGIs of the housekeeping genes that are consistently unmethylated (according to literature, the mPod data set and our computational predictions) to form the negative data set. Finally, we build classifiers to detect the CGIs with high potential of differential methylation patterns in cancerous conditions.

The CGI map for the study is obtained by applying the traditional Gardiner-Garden sequence criteria on non-repetitive sequences of the human genome [84].

The core steps of our model development to analyze CGI methylation status and aberrant methylation potential of human genome consist of three parts - feature extraction, feature selection and model construction through prediction tests.

For feature extraction, we use various resources to obtain or calculate both genetic features (CGI specific attributes, DNA compositional attributes, structural attributes, gene functional attributes, attributes related to evolutionary conservation, and attributes related to transcription factor binding) and epigenetic features (histone methylation and histone acetylation).

For feature selection, an automatic statistical test selection algorithm (section 5.2.1) is

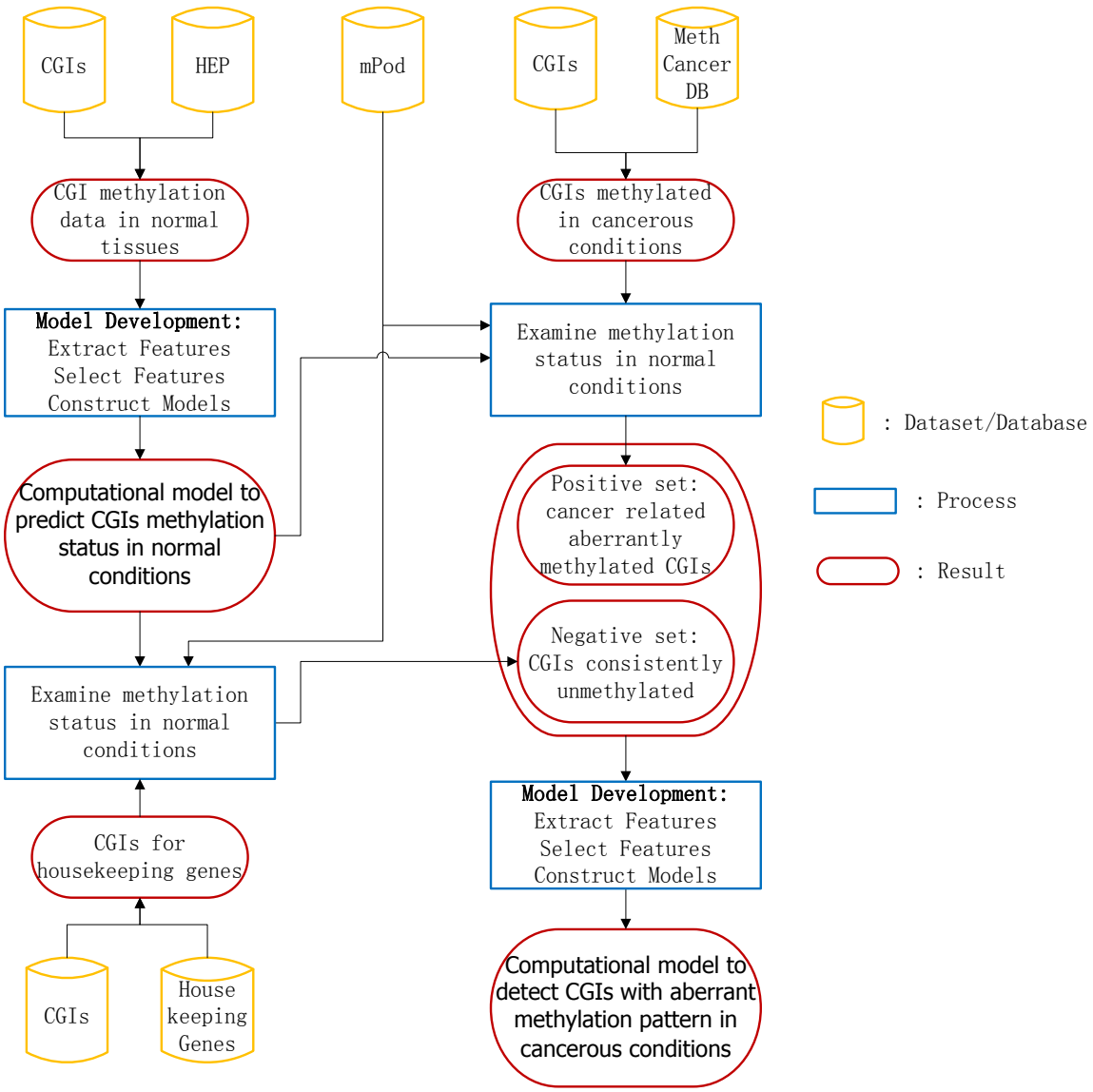

Figure 1: Workflow used for the prediction of methylation status and cancer related aberrant methylation of CGIs in human genome.

developed to identify features whose association with the methylation and aberrant methylation of CGIs is statistically significant. We then use PCA to decorrelate the selected features to further reduce dimensionality.

Finally, supervised learning based prediction tests are carried out to demonstrate to what degree the CGI methylation status and aberrant methylation potential of the human chromosomes can be identified using those principal components. The training and modified

cross-validation in this study are performed using various data sets including HEP, mPod, and MethCancerDB, and the assignment of aberrant methylation potential of CGIs are based on the trained predictive models. These assignments are performed on all promoter CGIs in the genome to prioritize potential novel aberrantly methylated genes. We give a detailed description of each step in Chapters 4 and 5.

CHAPTER IV

TRAINING DATA

We prepare training data for constructing CGI methylation predictive models in normal tissues, and the aberrant methylation predictive models in cancer, respectively. The traditional Gardiner-Garden criteria for CGI is adopted: (i) with $\geq$200 base pairs (bps), (ii) with a GC content >50%, and (iii) with an observed/expected CpG ratio $\geq$60% [5]. Based on these criteria, we obtain 27,639 CpG islands from the University of California, Santa Cruz (UCSC) Human Genome Browser [85]. We incorporate various DNA methylation and aberrant methylation resources including the HEP data set, mPod data set, MethCancerDB database, and various databases such as Cancer Genes [86] for functional annotation of cancer related genes.

4.1 Training Data for Methylation Prediction in Normal Tissues

High-resolution methylation profiles of human chromosomes 6, 20, 22 are obtained from the HEP data set [54]. The HEP data set provides a resource of about 1.9 million CpG methylation values of 2,524 amplicons derived from 12 different tissues in 43 samples using bisulfite DNA sequencing. The 12 different tissues consists of heart muscle, skeletal muscle, liver, sperm, fetal skeletal muscle, fetal liver, placenta, melanocytes, dermal fibroblasts, dermal keratinocytes, CD8 lymphocytes, and CD4 lymphocytes.

The HEP methylation intensity data of the CpG dinucleotides is calculated by comparing the C to T peaks at CpG sites [53]. The methylation intensity value of the analyzed CpGs ranges from 0 to 100 inclusive, where a value of zero corresponds to the lowest methylation intensity and a value of 100 to the highest methylation intensity. The genomic coordinates of the HEP data set are based on an old version of human genome assembly, and we map the coordinates of these CpG dinucleotides to the human genome assembly NCBI36/hg18 using the UCSC Genome Browser liftOver tool [84]. We extract the CGIs more than 10% of whose CpG dinucleotides are annotated with methylation intensities, and those CGIs

constitute our training data set for constructing the model to predict CGI methylation. For each tissue, the methylation intensity of a particular CpG dinucleotide was calculated as the average of the same sites detected from different samples [87]. Then the methylation intensity of a certain CGI is obtained by averaging over all the detected CpG dinucleotides within it. CGIs with methylation intensity greater than 50 were regarded as the methylated group, while less than 10 are treated as the unmethylated group [79].

Altogether, we have obtained 368 unmethylated CGIs and 101 methylated CGIs from HEP CD4 lymphocytes for training our methylation predictive models. For generalizability tests of the predictive models, we also extract the methylated and unmethylated CGIs in other 11 tissue and cell types from the HEP data set. The distribution of CGIs of both categories in each tissue or cell type is summarized in Table 3.

Table 3: Distribution of methylated and unmethylated CGIs among twelve different tissue and cell types from chromosomes 6, 20, and 22 based on the CpG methylation data from HEP.

Tissue/Cell type	**Methylated CGI**	**Unmethylated CGI**
CD4	101	368
CD8	103	332
sperm	45	331
liver	105	334
heart muscle	96	372
skeletal muscle	91	371
fetal skeletal muscle	79	281
fetal liver	76	270
placenta	92	328
dermal melanocytes	107	326
dermal fibroblasts	92	358
dermal keratinocytes	91	374

4.2 Training Data for Cancer Related Aberrant Methylation Prediction

For cancer related aberrant methylation prediction, our training data set consists of cancer related aberrant methylated CGIs (positive) and the CGIs that are consistently unmethylated (negative). Figure 2 illustrates how we incorporate different information resources to obtain the positive and negative groups.

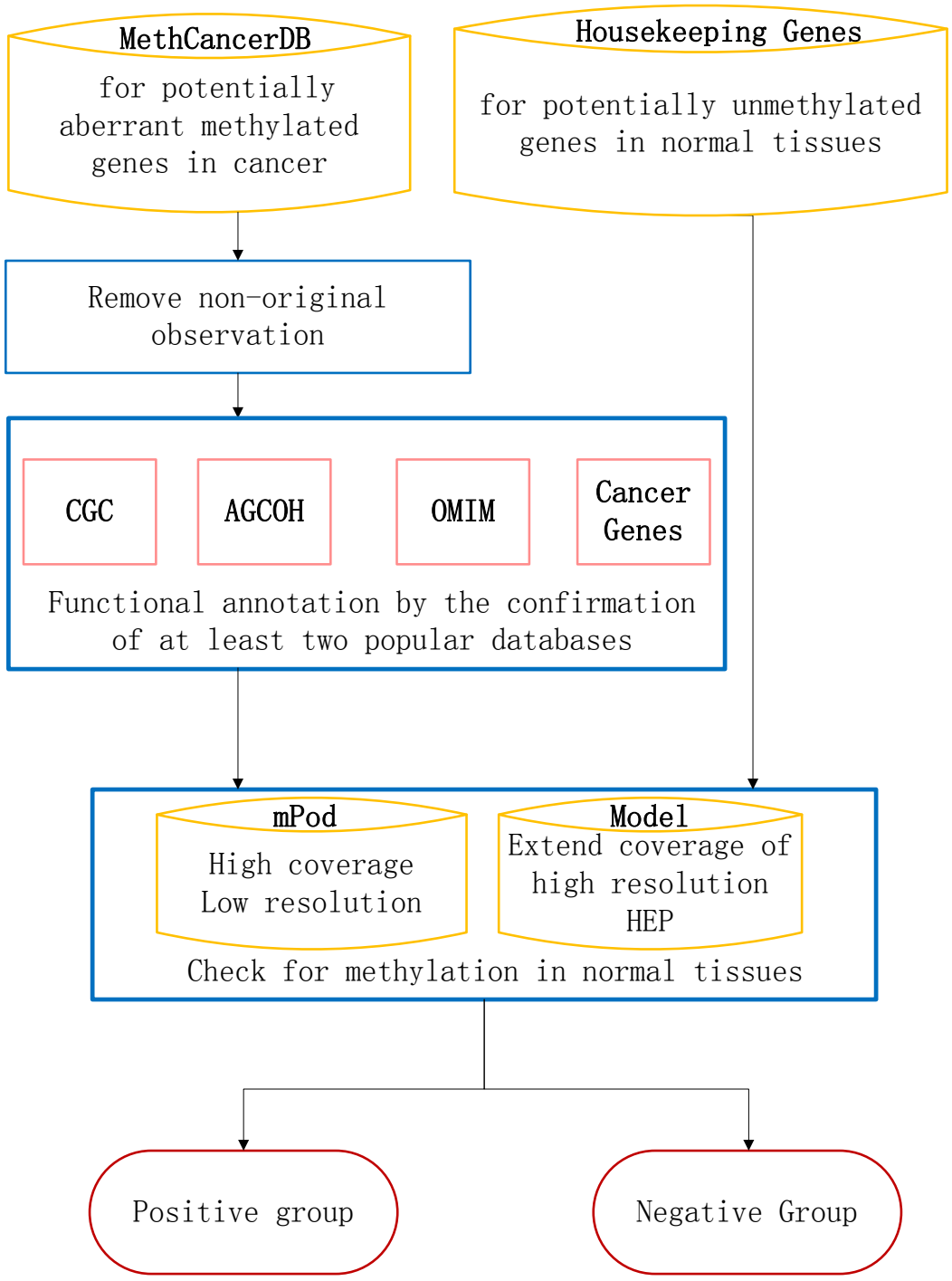

Figure 2: Quality control and functional annotation incorporating various resources for cancer related aberrantly methylated CGIs (positive) and the CGIs that are consistently unmethylated (negative).

The positive data set is formed by integrating five databases, including MethCancerDB [64], Cancer Gene Census (CGC) [88], Atlas of Genetics and Cytogenetics in Oncology and Haematology (AGCOH) [89], Online Mendelian Inheritance in Man (OMIM) [90], and Cancer Genes [86], that specify the (genetic vs. epigenetic) nature of cancer-related genes. We only consider those genes that are original observations of aberrant methylation in cancer from the MethCancerDB database, and extract the CGIs in their promoter regions. Since the MethCancerDB database does not specify the functional type of aberrantly methylated genes (e.g. oncogene or TSG), we further annotate the functional type of these genes by the support from at least two of the popular cancer databases, including CGC, AGCOH, OMIM, and Cancer Genes.

The negative data sets are formed by using the genes provided in [91] that are consistently expressed at high levels across 42 different tissues and cell types. These genes are named housekeeping genes and are believed to constitute a small set of genes required to maintain minimum basic cellular function [91]. We treat these genes as the negative control group based on the hypothesis that the promoter regions of the highly expressed genes are generally unmethylated.

We observe, however, that so-generated data sets may contain errors. For example, the gene KCNE4 (potassium voltage-gated channel, Isk-relate family, member 4) is with aberrant methylation according to the MethCancerDB database, but has been reported in [92] as methylated in normal tissues/cell types. To select the CGIs that are truly with cancer related aberrant methylation and consistently unmethylated in normal tissues, we used two more information resources, the mPod data set [47][48] and our computational predictive model for methylation status built on the HEP data set (section 4.1) to provide genome-wide prediction of the methylation status of CGIs in normal tissue/cell types, for further processing [93].

We select mPod because of its high coverage of the genome that can complement the HEP data set, and the diversity of tissue and cell types. The mPod data set contains the genome-wide DNA methylation profiles from 16 normal tissues/cell types that were obtained by using the MeDIP-chip technology accompanied with bioinformatics processing.

It has a relatively high coverage (as shown in Table 4) that characterizes the methylation status of 69,510 genomic fragments, but a relatively low resolution level in that each of these genomic fragment is ∼500 bps long. In contrast, the HEP data set, which contains the DNA methylation profiles of 12 normal tissues/cell types and is based on bisulfite sequencing technology, has a relatively high resolution level that specifies the methylation status of CpG dinucleotides but a relatively low coverage that only covers 1.9 million CpG dinucleotides (corresponding to 553 CGIs with ≥10% of whose CpG dinucleotides are covered) of three chromosomes (chr6, chr20, and chr22). To utilize both the mPod and HEP data sets to check the methylation status of those potentially aberrantly methylated and consistently unmethylated CGIs in the normal tissues and cell types, we first extend the coverage of the HEP data set by developing a computational predictive model that used the HEP data set for training and validation, and then use the mPod data set and this computational model to select those CGIs that are consistently identified as unmethylated in normal tissues. It is worthy of being pointed out that this extension and combination is meaningful because (i) the computational predictive model is faithful to the HEP data set (accuracy of ∼93-94% , specificity of ∼94% , and sensitivity of ∼92-93%), and (ii) the mPod data set and predictions from the computational model are not identical but highly correlated (correlation coefficient = 0.84).

The coordinates of the genomic regions in the mPod data set are based on the human genome assembly NCBI build 36/hg18. Each genomic region in this data set typically contains 5×50-mer probes. For each genomic region, the methylation intensity in a certain tissue is averaged over the methylation intensities of the probes within it [94]. Then the methylation intensity of a certain CGI is calculated by averaging over all the 500 bp ROIs overlapping with it.

Altogether, we identify 78 cancer related aberrantly methylated TSGs despite the cancer type, 177 cancer related aberrantly methylated genes for colon cancer, and 122 cancer related aberrantly methylated genes for prostate cancer as the positive groups. We also select a set of 783 housekeeping genes that are consistently unmethylated as the negative control group. The genomic locations for the aberrantly methylated TSG genes are summarized in

Table 5, which includes the gene symbol, chromosome number, start and end position, as well as strand information.

Table 4: CGI coverage of mPod for the methylation status in each chromosome. The number in parentheses represents the total number of CGIs in the corresponding chromosome. BC: B-cells, CX: cervix, CN: colon, LR: liver, LG: lung, PL: placenta, RM: rectum, PS: pancreas, PR: prostate, SM: skeletal muscle, SP: sperm, US: uterus, WB: whole blood.

Tissue	**chr1** (2463)	**chr2** (1680)	**chr3** (1159)	**chr4** (1019)	**chr5** (1227)	**chr6** (1251)	**chr7** (1552)	**chr8** (1028)	**chr9** (1230)	**chr10** (1150)	**chr11** (1371)
'BC'	2259	1530	1131	938	1141	1044	1325	900	1072	1037	1313
'CD4'	2259	1530	1131	938	1141	1044	1325	900	1072	1037	1313
'CD8'	2259	1530	1131	938	1141	1044	1325	900	1072	1037	1313
'CN'	2259	1530	1131	938	1141	1044	1325	900	1072	1037	1313
'CX'	2259	1530	1131	938	1141	1044	1325	900	1072	1037	1313
'GM'	2258	1530	1131	938	1141	1044	1325	900	1072	1037	1313
'LG'	2259	1530	1131	938	1141	1044	1325	900	1072	1037	1313
'LR'	2259	1530	1131	938	1141	1044	1325	900	1072	1037	1313
'PL'	2259	1530	1131	938	1141	1044	1325	900	1072	1037	1313
'PR'	2259	1530	1131	938	1141	1044	1325	900	1072	1037	1313
'PS'	2259	1530	1131	938	1141	1044	1325	900	1072	1037	1313
'RM'	2259	1530	1131	938	1141	1044	1325	900	1072	1037	1313
'SM'	2259	1530	1131	938	1141	1044	1325	900	1072	1037	1313
'SP'	2259	1530	1131	938	1141	1044	1325	900	1072	1037	1313
'US'	2259	1528	1131	938	1141	1043	1325	900	1070	1037	1312
'WB'	2259	1530	1131	938	1141	1044	1325	900	1072	1037	1313

Tissue	**chr12** (1221)	**chr13** (605)	**chr14** (788)	**chr15** (787)	**chr16** (1491)	**chr17** (1622)	**chr18** (508)	**chr19** (2544)	**chr20** (799)	**chr21** (356)	**chr22** (716)
'BC'	1159	533	738	712	1305	1488	466	2382	754	313	640
'CD4'	1159	533	738	712	1305	1488	466	2382	754	313	640
'CD8'	1159	533	738	712	1305	1488	466	2382	754	313	640
'CN'	1159	533	738	712	1305	1488	466	2382	754	313	640
'CX'	1159	533	738	712	1305	1488	466	2382	754	313	640
'GM'	1159	533	738	712	1305	1488	466	2382	754	313	640
'LG'	1159	533	738	712	1305	1488	466	2382	754	313	640
'LR'	1159	533	738	712	1305	1488	466	2382	754	313	640
'PL'	1159	533	738	712	1305	1488	466	2382	754	313	640
'PR'	1159	533	738	712	1305	1488	466	2382	754	313	640
'PS'	1159	533	738	712	1305	1488	466	2382	754	313	640
'RM'	1159	533	738	712	1305	1488	466	2382	754	313	640
'SM'	1159	533	738	712	1305	1488	466	2382	754	313	640
'SP'	1159	533	738	712	1305	1488	466	2382	754	313	640
'US'	1159	531	738	710	1305	1488	466	2382	754	313	640
'WB'	1159	533	738	712	1305	1488	466	2382	754	313	640

Table 5: Genomic location for the 78 cancer related aberrantly methylated TSGs despite the cancer type.

Gene Symbol	Chromosome Number	Start	End	Strand
AKR1B1	chr7	133777646	133794428	-
APAF1	chr12	97563208	97653342	+
APC	chr5	112101454	112209835	+
APP	chr21	26174731	26465003	-
BRCA2	chr13	31787616	31871809	+
C2orf40	chr2	106048544	106061041	+
CAMTA1	chr1	6767970	7752351	+
CASP8	chr2	201830998	201860679	+
CAV1	chr7	115952074	115988466	+
CD82	chr11	44543716	44597891	+
CDH1	chr16	67328695	67426945	+
CDH13	chr16	81218078	82387700	+
CDKN1B	chr12	12761575	12766570	+
CDKN1C	chr11	2861023	2863571	-
CFLAR	chr2	201689060	201737248	+
CNTN4	chr3	2117246	3074645	+
CXCR4	chr2	136588388	136592195	-
CYP27B1	chr12	56442383	56447243	-
DKK1	chr10	53744046	53747423	+
DKK3	chr11	11941118	11986762	-
DLC1	chr8	12985242	13416766	-
DPP4	chr2	162557000	162639298	-
EFNA5	chr5	106744249	107034495	-
FABP3	chr1	31610686	31618510	-
FAS	chr10	90740267	90765522	+
FHIT	chr3	59710075	61212173	-
GJB2	chr13	19659604	19665114	-
GNMT	chr6	43036477	43039596	+
GPX3	chr5	150380191	150388747	+
GSTP1	chr11	67107861	67110699	+
HIC1	chr17	1906353	1909731	+
IGFBP7	chr4	57592000	57671296	-
ING4	chr12	6629964	6642569	-
INTS6	chr13	50833701	50925276	-
IRF8	chr16	84490274	84513712	+
KL	chr13	32488570	32538279	+
KLK10	chr19	56207811	56215243	-
LATS1	chr6	150023743	150081085	-
LOX	chr5	121429917	121441853	-
MLH1	chr3	37009982	37067341	+
MSH2	chr2	47483766	47563864	+
MTHFR	chr1	11768373	11788702	-
PGR	chr11	100405564	100505754	-
POU2F3	chr11	119616160	119695863	+
PPP1R1B	chr17	35036704	35046404	+
PRDM2	chr1	13903936	14024162	+
PRKCDBP	chr11	6296751	6298316	-
PTEN	chr10	89613174	89718512	+
PTGS2	chr1	184907591	184916179	-
PTPN13	chr4	87734908	87955326	+
PTPRG	chr3	61522284	62254738	+
PTPRO	chr12	15366753	15641602	+
RARB	chr3	25444757	25614424	+
RASSF1	chr3	50342220	50353371	-
RASSF2	chr20	4708668	4743769	-
RASSF5	chr1	204747501	204829239	+
RB1	chr13	47775883	47954027	+
S100A2	chr1	151800208	151804930	-
SERPINB5	chr18	59295123	59323298	+
SFN	chr1	27062219	27063534	+
SFRP1	chr8	41238634	41286137	-
SFRP5	chr10	99516497	99521746	-
SLIT2	chr4	19864332	20229886	+
SOCS1	chr16	11255774	11257540	-
SOCS3	chr17	73864456	73867753	-
SPRY2	chr13	79808112	79813087	-
STK11	chr19	1156797	1179434	+
TES	chr7	115637816	115686073	+
TNFRSF10A	chr8	23104914	23138584	-
TP53	chr17	7512444	7531588	-
TPM1	chr15	61121890	61151166	+
TSC22D1	chr13	43905654	44048701	-
UCHL1	chr4	40953685	40965203	+
VHL	chr3	10158318	10168746	+
WIF1	chr12	63730672	63801383	-
WRN	chr8	31010319	31150819	+
XAF1	chr17	6599879	6619688	+
ZMYND10	chr3	50353540	50358160	-

CHAPTER V

METHODS

5.1 Feature Extraction

As described in section 2.4.3, we investigate the association between CGI methylation and eight categories of related feature classes. It is worth being pointed out that, among the eight categories, we incorporate three sets of features that have not been extensively explored previously, including (*i*) the nucleosome positioning propensities of the CGI, (*ii*) the acetylation status of nearby histones, and (*iii*) the functional roles of nearby genes. These features add more dimensions of information as shown by PCA in section 6.1.1.

In the following paragraphs of A.1 to A.8., we describe how the features in these eight categories are extracted.

A.1. The CGI specific attributes, including the GC content, length, and observed/expected CpG ratio, are directly obtained from the UCSC human genome browser.

A.2. For the DNA compositional features, we focus on the frequencies of the tetramer oligonucleotides and their z-scores; The z-score of a tetramer oligonucleotide fragment, $N_1N_2N_3N_4$, is calculated as:

$$Z(N_1N_2N_3N_4) = \frac{O(N_1N_2N_3N_4) - E(N_1N_2N_3N_4)}{\sigma(N_1N_2N_3N_4)}, \tag{4}$$

where $O(\cdot)$ represents the observed frequency, $E(\cdot)$ and $\sigma(\cdot)$ represent the expected frequency and standard deviation. $E(N_1N_2N_3N_4)$ was estimated empirically based on a maximal-order Markov model [95]:

$$E(N_1N_2N_3N_4) = \frac{O(N_1N_2N_3)O(N_2N_3N_4)}{O(N_2N_3)}, \tag{5}$$

and $\sigma(N_1N_2N_3N_4)$ was approximated as:

$$\begin{aligned}\sigma(N_1N_2N_3N_4) &= E(N_1N_2N_3N_4)* \\ &\frac{[O(N_2N_3) - O(N_1N_2N_3)][O(N_2N_3) - O(N_2N_3N_4)]}{O^2(N_2N_3)}.\end{aligned} \tag{6}$$

A.3. For the DNA structural features, we focus on those basic characteristics capturing the DNA 3-D conformation as well as the nucleosome positioning propensities. The DNA conformation related attributes include twist, tilt, roll, shift, slide and rise, which are estimated based on a model of dinucleotide stiffness [96]. For each of these six attributes, the average value over all dinucleotides of the CGI is calculated.

Nucleosome positioning propensities of the CGIs are estimated based on the genome-wide prediction of the nucleosome organization map [97]. There are two types of predictions, one at the nucleotide level, and the other at the DNA fragment level. The nucleotide level prediction regards the probability of each nucleotide being covered by any nucleosome, which we calculate based on the mean and standard deviation over the entire CGI. The fragment level prediction regards the nucleosome positioning potential of each 147 bp (the typical length of a nucleosome) DNA fragment, which we calculate based on the mean and standard deviation over all fragments overlapping with the CGI.

A.4. For the distribution of TFBSs, we download the data set for the locations and scores of TFBS conserved in the human/mouse/rat alignment from the UCSC Genome Browser [84]. A binding site can be considered to be conserved across the alignment if its score meets the threshold score for its binding matrix in all 3 species [84]. Altogether, we obtain 115 TFBSs and map their genomic location to the CGIs. We then extract the features of the occurrences and mean scores of each TFBS associated with each CGI. Furthermore, due to the observed association of TFBS with CGI flanking regions [98], we extend both sides of the CGIs up to 2,000 bp with a step size of 100 bp, and then recalculate the occurrences and mean scores of each TFBS overlapping with the flanked CGIs.

A.5. We download the predictions of conserved elements from the UCSC Genome Browser [84]. These predictions by the phastCons [82] provide evolutionarily conserved elements across vertebrate, insect, worm, and yeast genomes. Each element is associated with a log-odds score quantifying its degree of conservativeness across genomes. The score is equal to its log probability under the conserved model minus its log probability under the non-conserved model [84]. We retrieve the genomic coordinates for the conserved sites from the table named phastConsElements17way, and then extract the number and mean scores

of the evolutionary conserved elements overlapping with each CGI. To account for both the short- and long-range association between these elements and CGIs, we consider flanking regions of various lengths, ranging from 100 bps to 2,000 bps (with step size of 100 bps) upstream and downstream of the CGI. We count the number of these elements overlapping with the CGI, and calculated their average score.

A.6. We examine whether a CGI's nearby genes are involved in any cancer-related biological processes. A CGI's nearby genes refer to those whose promoter region (from the 1,000 bps upstream to the 200 bps downstream of the transcription start site) overlaps with the CGI. A total of 37 biological processes (30 oncogene related, 11 tumor suppressor related, and 4 common) are determined through gene ontology enrichment analysis of the genes retrieved from the Cancer Gene Census [99]. A cancer-related biological process was considered to be enriched if (*i*) the number of genes involved in the process is larger than five, and (*ii*) the enrichment factor of the process is greater than one. If the gene ontology annotations of a gene include one or more of these processes, the corresponding gene function feature is assigned a value of 1, and 0 otherwise.

A.7. The histone methylation information is obtained from Barski, *et al.*'s data set, which characterizes the genome wide distribution of 20 histone methylations, as well as histone variants H2A.Z, RNA polymerase II, and the insulator binding protein CTCF in CD4 lymphocytes [100]. Altogether we obtain about 186.9 million sequence tags, and the number of tags detected for a particular position can be treated as proportional to the histone methylation level of that position. We use the mean and standard deviation of the number of tags over all nucleotides of a CGI to represent the methylation level of the CGI's nearby histones.

A.8. We obtain about 88.9 million sequence tags for 18 histone acetylations in CD4 lymphocytes from Wang, *et al.*'s data set [101]. These 18 histone acetylations are H2AK5ac, H2AK9ac, H2BK5ac, H2BK12ac, H2BK20ac, H2BK120ac, H3K4ac, H3K9ac, H3K14ac, H3K18ac, H3K23ac, H3K27ac, H3K36ac, H4K5ac, H4K8ac, H4K12ac, H4K16ac, and H4K91ac. We then extract the genomic coordinates of these sequence tags, and map the obtained tags of histone acetylation to the CGIs. We score the histone acetylation modification of each

nucleotide position within the CGI by the number of tags covering that position. We then calculate the mean and standard deviation of the histone acetylation intensities across all the nucleotides within each CGI.

5.2 Feature Selection

The raw feature dimension is very high and the performance of supervised learning methods would suffer from the curse of dimensionality. Aspects of the curse of dimensionality include the need of large volume of training data to achieve high generalization accuracy, and the time complexity of training the supervised learning methods [102]. As a result, selection of subsets from the extracted large number of genomic and epigenetic features is important for the predictive model construction. We thus carry out a two-stage selection procedure: statistical test followed by PCA.

5.2.1 Statistical Test

There are three candidate statistical tests evaluated in this study to identify features whose association with the methylation status of CGIs is statistically significant. The three statistical tests involved are as follows.

1) Fisher's exact test [103], which is a statistical significance test to determine whether there are nonrandom associations between two nominal variables. It is often used in the analysis of contingency tables where sample sizes are small.

2) Chi-squared test [104], which is to determine if there is a significant difference between the expected values and the observed values in one or more categories. Yates' correction [105] can be incorporated to improve the mathematical approximation for the test statistics.

3) Kolmogorov-Smirnov (KS) test [106], which compares the distributions of the values in two samples to determine if they are from the same continuous distribution. The KS test is distribution-free (i.e. makes no assumption about the distribution of data).

We implemented an algorithm to automatically select an appropriate statistical test for the association analysis between the extracted features and the methylation status of CGIs. The algorithm selects the optimal statistical test based on the following criteria. When the feature variable is continuous, the algorithm selects the KS test. When the feature variable

is categorical, the algorithm decide whether any of the expected values in the contingency tables is extremely small (<5). If yes, the algorithm selects the Fisher's exact test; otherwise, the algorithm selects the Chi-squared test with Yates' correction incorporated. The readers are referred to [107][108] for the underlying mathematical principal for the statistical test selection. For each of the tests, a feature is considered to be statistically significantly associated with the methylation status of CGIs if the p-value yielded by the test is less than 0.05.

5.2.2 PCA

Besides their correlations with the CGI methylation or aberrant methylation, the identified features might be inter-correlated. For example, the histone methylation and acetylation status are likely to be correlated, because some acetylation and methylation (e.g. histone H3 at lysine 9) play opposite roles in gene activity [109]; DNA sequence and structure properties are likely to be correlated, because most DNA structures are predicted based on DNA sequences; and, the distribution of functional/evolutionarily conserved elements in a short flanking neighborhood (e.g., +/- 200 bps) is likely to be correlated with the distribution in a longer flanking neighborhood (e.g., +/- 2000 bps). The correlation between features makes the feature space unnecessarily high-dimensional. To minimize the redundancy in the features, we perform the PCA on those methylation-related features that are selected via the above statistical tests. The PCA uses an orthogonal transformation to convert a set of values of possibly correlated dimensions into a set of values of uncorrelated dimensions called principal components [110]. Technically, given a collection of vectors in the original feature space, $\{\mathbf{x}\} \subset X$, the mean and covariance matrix of $\{\mathbf{x}\}$ are denoted as:

$$\boldsymbol{\mu}_{\mathbf{x}} = E\{\mathbf{x}\} \tag{7}$$

$$\mathcal{C}_{\mathbf{x}} = E\{(\mathbf{x} - \boldsymbol{\mu}_{\mathbf{x}})(\mathbf{x} - \boldsymbol{\mu}_{\mathbf{x}})^T\} \tag{8}$$

where $E\{\cdot\}$ represents the expected value. The components of $\mathcal{C}_{\mathbf{x}}$, denoted by c_{ij}, represents the covariance between the i^{th} and j^{th} feature components. Via the singular value decomposition [111] of the covariance matrix $\mathcal{C}_{\mathbf{x}}$, PCA derives a unitary matrix, $\Theta \in \mathbb{R}^{M \times M}$, to map each $\mathbf{x}$ in the original M-dimensional feature space X to $\mathbf{y} \equiv [y_1, y_2, ..., y_M]^T$ in a new M-dimensional feature space $Y \subset \mathbb{R}^M$ via the linear transformation

$$\mathbf{y} = \Theta(\mathbf{x} - \boldsymbol{\mu}_{\mathbf{x}}) \tag{9}$$

Note that the order of new feature elements, y_1, y_2,..., y_M, is related to the magnitude of the projection of $\mathcal{C}_{\mathbf{x}}$ onto the i^{th} $(i = 1, ..., M)$ axis in the new coordinate system [112]. That is, the first dimension (y_1) corresponds to the projection with the largest magnitude, the second (y_2) corresponds to the projection with the second largest magnitude, etc. This means that after PCA transformation, the feature components are completely decorrelated, and the information contained in the original feature space before the transformation is maximally retained in the first several number of components of the new feature space. Therefore, by keeping only the first several components of the new feature space, most of the information can still be retained while the redundancy in the feature collection is greatly removed and the dimensionality of the feature space is greatly reduced. For PCA to work properly, we subtract the mean from each of the feature dimensions.

5.3 Control Group

Housekeeping genes generally refer to constitutive genes that are necessary for the maintenance of basic cellular function [113]. Since housekeeping genes constitutively expressed in all tissues and their expression levels are comparatively constant across different cell types, the corresponding CGIs are generally considered to be unmethylated [91]. To construct the predictive model for the detection of the genes that are subject to aberrant methylation in cancer, we selected a set of 783 housekeeping genes as the negative control for the cancer-related aberrant methylated genes [91]. These housekeeping genes are consistently expressed as demonstrated by the gene expression profiles of 42 normal human tissues on custom high density microarrays [91]. And, these housekeeping genes are further demonstrated to be constitutively unmethylated across normal tissues as confirmed by the mPod

data set and our computational predictive model built from HEP data set. Furthermore, these housekeeping genes do not overlap with any of the aberrantly methylated genes in colon and prostate cancers as shown by the MethCancerDB. We mapped the genomic coordinates of these housekeeping genes to the NCBI36/hg18 assembly through UCSC Genome Browser [114].

5.4 Prediction Test

We perform the support vector machine-based prediction test in three broad scenarios: DNA methylation status binary classification, DNA methylation regression, and cancer-related aberrant methylation detection. For DNA methylation status classification, we build computational models to identify whether a CGI is methylated or unmethylated in normal tissues; For DNA methylation regression, we predict the methylation level (continuous variable) of each CGI; And for cancer-related aberrant methylation detection, we identify CGIs that are potentially aberrantly methylated in cancer, including colon and prostate cancer.

In all scenarios, the inputs for the classifier construction are normalized between [0,1]. Since the accuracy of a support vector machine model is largely dependent on the selection of the model parameters, we need to find the optimal parameter values in the (C,γ) space, where C is the regularization parameter, and γ is the kernel parameter for the radial basis function. We perform a two-stage grid search approach [115] in each fold of the cross-validation experiment on the training data. In the first stage, we use a coarse grid with 10 search regions for each parameter. After identifying a better region on the grid, a finer grid search on that region is conducted. The optimal values of (C,γ) pair are determined when they yield the largest classification accuracy [116].

5.4.1 DNA Methylation Status Classification

The features selected through statistical tests and PCA are used to build support vector machine-based models to predict the CGI methylation status. To examine the contribution of the newly added features as well as the impact of the inhibitive-to-acquire histone modification information, we establish the following predictive models, (1) M_1: a model with all information being incorporated, (2) M_2: a model with all but the histone modification

information being incorporated, (3) M_3–M_9: seven models with individual or combinations of the newly added features being excluded, and (4) M_{10}–M_{16}: seven models with individual or combinations of the newly added features as well as the histone methylation information being excluded. We use the CD4 lymphocyte data for training and validating the models, while the data of the other 11 tissues/cell types for generalizability testing.

All these models are trained and validated by using a 10-fold cross validation scheme. That is, all CGIs are partitioned randomly into 10 approximately equally-sized folds, each of which is used in turn for validation while the remaining folds are used for training. The performance of the classifiers is assessed by using three metrics defined in Eqns. (2)–(3), namely, sensitivity (SE), specificity (SP), and accuracy (ACC). This partition-training-and-validation procedure is repeated for 20 times, and the classifier performance is averaged over the 200 validation folds. For fair comparisons with the existing method, a leave-one-out cross-validation (LOOCV) scheme is also used. That is, each CGI is in turn used for validation while the remaining CGIs are used for training. The performance of the model in the LOOCV scheme is also assessed by the three metrics averaged over all validation CGIs.

Two predictive models built on the CD4 lymphocyte data are tested for generalizability using the data of the other 11 tissues and cell types: one (M_1) relying on all information, while the other (M_2) relying on all but the histone modification information. For the former model, because the genome-wide histone methylation and acetylation profiles are not available for these 11 tissues and cell types, we use the genome-wide histone modification profiles in the CD4 lymphocytes, assuming that histone modifications in various cell types are moderately or even highly correlated [117]. We also calculate the Pearson product moment correlation coefficients of the CGI methylation levels across different tissues and cell types to further support our computational results. Pearson product moment correlation coefficient between two random variables X and Y is defined as

$$\mathbf{r} = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{n(\Sigma x^2 - (\Sigma x)^2}\sqrt{n(\Sigma y^2 - (\Sigma y)^2}} \tag{10}$$

The value of r is such that $-1 < r < +1$. The $+$ and $-$ signs are used for positive linear correlations and negative linear correlations, respectively.

5.4.2 DNA Methylation Regression

To profile the specific methylation level of a CGI, we construct a regression model to predict the profile of the methylation levels of the CGI using the CD4 lymphocyte data. Like the binary classification model, the 10-fold cross validation scheme is implemented for the training and validation, and is repeated 20 times. However, unlike the binary classification model, a support vector regression model is designed to generate values ranging from 0 to 100, with 0 representing the lowest methylation intensity and 100 representing the highest methylation intensity. The methylation intensity values are averaged over all the validations and repeats to filter noise. We also perform the correlation analysis between the methylation intensity generated by the regression models and the methylation value calculated by the HEP data set.

5.4.3 Cancer-related Aberrant Methylation Prediction

The aberrant methylation of CGIs has been ascribed to the onset and progression of human cancers [81][118]. We take advantage of the features selected through statistical tests and PCA to construct support vector machine-based models to predict the CGIs that are subject to aberrant methylation in cancer. Histone modification features are not used in such prediction due to unavailability, and cancer-related gene function features are not used to eliminate any prior knowledge. We select the training genes for aberrant methylation in two different settings. In one setting, we select the aberrantly methylated TSGs, and in the other we select aberrantly methylated genes in a specific cancer type. Besides the CGIs for the aberrantly methylated genes, a set of 783 housekeeping genes are selected to serve as the negative control group to make comparisons.

The number CGIs in the negative control group greatly outnumber the aberrantly methylated genes in our training data. In this case, the traditional 10-fold cross validation has been found to be biased toward a group with the higher number of CGIs and could yield misleading outcome [119]. To account for the imbalance of the training data, we apply a modified 10-fold cross validation strategy to estimate the prediction accuracy. That is, the CGIs of the housekeeping genes are first randomly partitioned into non-overlapping

groups, with each group containing approximately the same number of CGIs as that of the aberrantly methylated genes. Then, for each non-overlapping group and all the aberrantly methylated genes, a stratified 10-fold cross validation is performed. That is, the aberrantly methylated group and the non-overlapping control group are each randomly partitioned into two portions. The portions with 10% of the CGIs from both groups are used for testing to estimate the prediction accuracy, while the other portions with 90% of the CGIs are used for building a support vector machine-based classifier. The above procedure is repeated for 20 times to obtain a reliable statistical estimate of the average prediction performance.

CHAPTER VI

RESULTS AND DISCUSSIONS

6.1 Methylation Status Classification

6.1.1 Statistical Tests and PCA

Out of a total number of 841 features, 342 features were retained whose p-values in the statistical tests were less than 0.05 [25, 93]. These features include two of the CGI specific attributes, 217 DNA compositional and eight DNA structural features, 35 functional element features and two evolutionarily conserved element features, two features regarding the functional roles of the neighboring genes, and 76 features related to the modification status of nearby histones. Particularly, among the newly added features, two out of the four nucleosome positioning features, all of the 36 histone acetylation features, and both of the features regarding the functional roles of the neighboring genes were retained after statistical tests.

Table 6: Number of principal components (PCs) required to retain a certain percentage (Pcnt) of the total variance.

Pcnt	100%	99.99%	99.90%	99.00%
PCs	342	10	8	6
Pcnt	95.00%	90.00	75.0%	50.00%
PCs	5	4	3	2

PCA was performed to decorrelate these 342 selected features. Table 6 summarizes the number of principal components that must be retained to keep a certain percentage of the variance of the original feature space. Observe that the first eight principal components together can account for the ∼99.90% of the variance in the original feature space and were therefore used to build the predictive models. Fig. 3 depicts the contribution of each of the 342 original feature dimensions to the eight principal components.

Observe from Fig. 3 that each of the following eight categories of features, (i) the

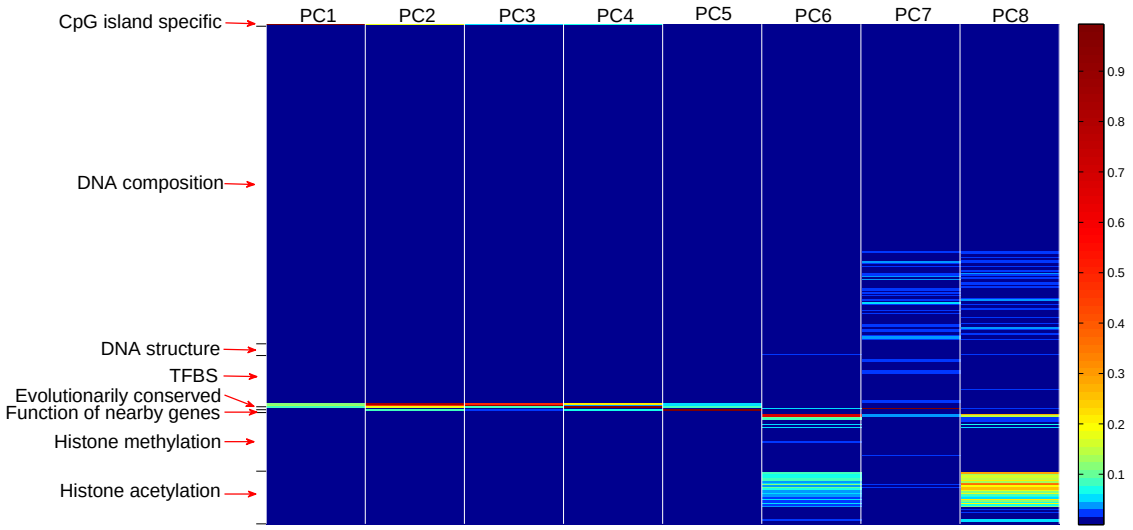

Figure 3: Contribution of the 342 features to the eight principal components. Each column corresponds to a principal component, and each row corresponds to an original feature dimension.

CGI specific attributes, (*ii*) DNA sequence patterns, (*iii*) DNA structure patterns, (*iv*) distribution of TFBS, (*v*) distribution of the evolutionarily conserved elements, (*vi*) gene functions, (*vii*) histone methylation and (*viii*) histone acetylation status, makes substantial contributions to one or more principal components, suggesting that these categories of information, though correlated, are complementary to a certain extent for predicting the CGI methylation.

6.1.2 Performance of the Predictive Models on the CD4 Lymphocyte

The specificity, sensitivity, and accuracy measures of our predictive model M_1 that incorporates all information are summarized in Table 7. Observe that both cross-validation schemes rendered similar results, indicating that these measures can reliably characterize our model. The performance of our classifier was compared to that of Fan et al.'s [79] method. Note that both models incorporated the histone modification information and were evaluated by using the HEP CD4 lymphocyte data to be fair. Observe that our model showed an improved specificity and accuracy than Fan, *et al.*'s model while maintaining a comparable sensitivity. Particular, our method achieved more than 20% boost in specificity in comparison with Fan, *et al.*'s model. Furthermore, it was reported in [79] that when evaluated on

the human brain data, Fan, *et. al.*'s method could outperform another DNA methylation prediction algorithm named Epigraph [77].

Table 7: Performance of our classifiers M_1 on CD4 lymphocytes with comparison to the existing method.

Method	**SP**	**SE**	**ACC**
M_1 (10-fold)	0.9405	0.9257	0.9313
M_1 (LOOCV)	0.9429	0.9307	0.9403
Fan et al.'s [79]	0.7400	0.9428	0.8994

The improvement of model M_1 over the existing model was partly due to the incorporation of the three new types of features – nucleosome positioning propensities, gene functions, and histone acetylation status. The performance of our models M_3 through M_9, each with an individual or a combination of the new types of features being excluded, are summarized in Table 8. Observe that the performance of the predictive model deteriorated to different extents when individual or combinations of the newly added features were excluded. Specifically, the models without histone acetylation information (M_3, M_6, M_7, and M_9) deteriorated more than those models with histone acetylation information but without the other two types of newly added features (M_4, M_5, and M_8). Therefore, histone acetylation appears to be the most influential feature to the performance of the predictive model among the newly added features.

We suspected that the information carried by the histone methylation features was too dominant to fairly assess the influence of these newly added features; and therefore excluded the histone methylation features and repeated the above experiments excluding individual or combinations of the newly added features. The resultant models were M_{10} through M_{16}, and their performance was summarized in Table 8. Similarly, the models without an individual or a combination of the newly added features deteriorated. It is noteworthy that (1) the histone methylation and acetylation information greatly affected the sensitivity of the models, and (2) the loss of histone methylation information could largely be made up by including the histone acetylation information. This is not surprising, given that these two forms of histone modifications are closely related as repeatedly observed in various tissues

Table 8: Statistical measurements for the performance of the predictive models (M_3 through M_{16}), each with an individual or a combination of the newly added categories of features being excluded.

	Features	SP		SE		ACC	
		LOOCV	10-fold	LOOCV	10-fold	LOOCV	10-fold
Histone Methyl Retained	All retained	0.9429	0.9405	0.9307	0.9257	0.9403	0.9313
	Acetylation (M_3)	0.9048	0.9012	0.9010	0.8965	0.9175	0.9046
	Function roles (M_4)	0.9319	0.9302	0.9315	0.9265	0.9362	0.9210
	Nucleosome (M_5)	0.9285	0.9270	0.9276	0.9250	0.9205	0.9205
	Acetylation + Function roles (M_6)	0.8876	0.8791	0.8912	0.8903	0.8915	0.8897
	Acetylation + Nucleosome (M_7)	0.8805	0.8698	0.8815	0.8835	0.8902	0.8826
	Function roles + Nucleosome (M_8)	0.9208	0.9186	0.9107	0.9116	0.9202	0.9186
	All three (M_9)	0.8775	0.8685	0.8810	0.8822	0.8806	0.8786
Histone Methyl Excluded	All but histone methylation	0.9321	0.9318	0.5941	0.5932	0.8593	0.8575
	Acetylation (M_{10})	0.9701	0.9670	0.2277	0.2247	0.8102	0.8001
	Function roles (M_{11})	0.9109	0.9092	0.5720	0.5670	0.8369	0.8312
	Nucleosome (M_{12})	0.9088	0.9078	0.5682	0.5660	0.8298	0.8296
	Acetylation + Function roles (M_{13})	0.9402	0.9320	0.2289	0.2279	0.7885	0.7862
	Acetylation + Nucleosome (M_{14})	0.9381	0.9266	0.2302	0.2304	0.7752	0.7641
	Function roles + Nucleosome (M_{15})	0.9012	0.8990	0.5520	0.5519	0.8252	0.8232
	All three (M_{16})	0.9098	0.8972	0.2341	0.2338	0.7406	0.7352

and cell types [109].

6.1.3 Classifier Generalizability

The two predictive models, one with the histone modification information (M_1) and the other without (M_2), that were both built on the human CD4 lymphocyte data were tested on the data of the other 11 tissue and cell types for their generalizability. The sensitivity, specificity, and accuracy of M_1 and M_2 during these testing experiments are summarized in Table 9 and Table 10, respectively.

When the histone modification information was incorporated, the classifier model constructed on the CD4 lymphocyte data can be applied to most of the other tissues and cell types (except for sperm) with little or no performance deterioration. When the histone modification information was not incorporated, the performance of the predictive model on the data of the other tissues and cell types deteriorated substantially, especially in terms

of the sensitivity. However, if compared to the validation results where the histone modification information was not used, the performance on the testing data was not unexpected. Therefore, with or without the histone modification information, the predictive model established on the CD4 lymphocyte data can well generalize to the other tissue or cell type data.

Table 9: Generalizability performance of the methylation predictive model constructed from the HEP CD4 lymphocytes data on 11 different tissues and cell types: with histone modification.

Procedure	**Tissue/Cell Type**	**SP**	**SE**	**ACC**
Validation	CD4 (10-fold)	0.9405	0.9257	0.9313
	CD4 (LOOCV)	0.9429	0.9307	0.9403
Testing	CD8	0.9608	0.8932	0.9448
	liver	0.9680	0.8762	0.9465
	heart muscle	0.9462	0.9479	0.9466
	skeletal muscle	0.9542	0.9451	0.9524
	embryonic skeletal	0.9395	0.9367	0.9389
	embryonic liver	0.9259	0.9342	0.9277
	placenta	0.9695	0.9130	0.9571
	dermal melanocytes	0.9663	0.8785	0.9446
	dermal fibroblasts	0.9525	0.9239	0.9467
	dermal keratinocytes	0.9385	0.9341	0.9376
	sperm	0.8459	0.9778	0.8617

Considering that DNA methylation is heavily involved in cellular differentiation, our results in Tables 9 and 10 look suspicious. We therefore calculated the correlations of the CGI methylation levels between different tissue and cell types, as depicted in Fig. 4. Observe that the correlation coefficients between the somatic/placenta cells are very high (mean: 0.9455, standard deviation: 0.0229), where the correlation coefficients between the somatic/placenta and sperm cells are only moderate (mean: 0.6706, standard deviation: 0.0225). This suggests that the methylation status of CGIs are highly correlated in various somatic/placenta cells, and therefore do not represent tissue-specific differentially methylated regions. Our observations are consistent with recent studies [46][120] that there are few variance in methylation levels of autosomal CGI promoters, and there is only a relatively small fraction of CGIs with tissue-specific methylation. The difference between the

Table 10: Generalizability performance of the methylation predictive model constructed from the HEP CD4 lymphocytes data on 11 different tissues and cell types: without histone modification.

Procedure	Tissue/Cell Type	SP	SE	ACC
Validation	CD4 (10-fold)	0.9670	0.2247	0.8001
	CD4 (LOOCV)	0.9701	0.2277	0.8102
Testing	CD8	0.9722	0.2108	0.8104
	liver	0.9678	0.2143	0.8122
	heart muscle	0.9562	0.2386	0.8186
	skeletal muscle	0.9594	0.2364	0.8306
	embryonic skeletal	0.9425	0.2298	0.8100
	embryonic liver	0.9389	0.2306	0.8054
	placenta	0.9655	0.2184	0.8276
	dermal melanocytes	0.9700	0.2186	0.8156
	dermal fibroblasts	0.9605	0.2200	0.8237
	dermal keratinocytes	0.9425	0.2204	0.8095
	sperm	0.8524	0.2365	0.7625

somatic/placenta and sperm cells, as reflected by their moderate cross-correlations and the performance deteriorations of our prediction models being applied to the sperm cell data, suggests that gametes are epigenetically more deviated from somatic cells than somatic cells themselves. This difference is likely related to the meiotic process, the special conditions and gene expression required for gamete production [121].

6.2 Methylation Status Regression

We further performed support vector regression on the HEP data set to construct a regression model which can generate the methylation intensity of CGIs. The methylation intensities predicted by the regression model for the 368 unmethylated CGIs and 101 methylated CGIs are illustrated in Fig. 5. The CGIs to the left of the red line are unmethylated from the HEP data set, and to the right are methylated. To validate the prediction, we performed Pearson correlation analysis of the predicted intensity with the intensity calculated from the HEP data set. The correlation analysis demonstrated that the predicted intensity is highly correlated with the intensity calculated from the HEP data set (correlation coefficient $\rho = 0.883$, p-value$<4\times10^{-122}$).

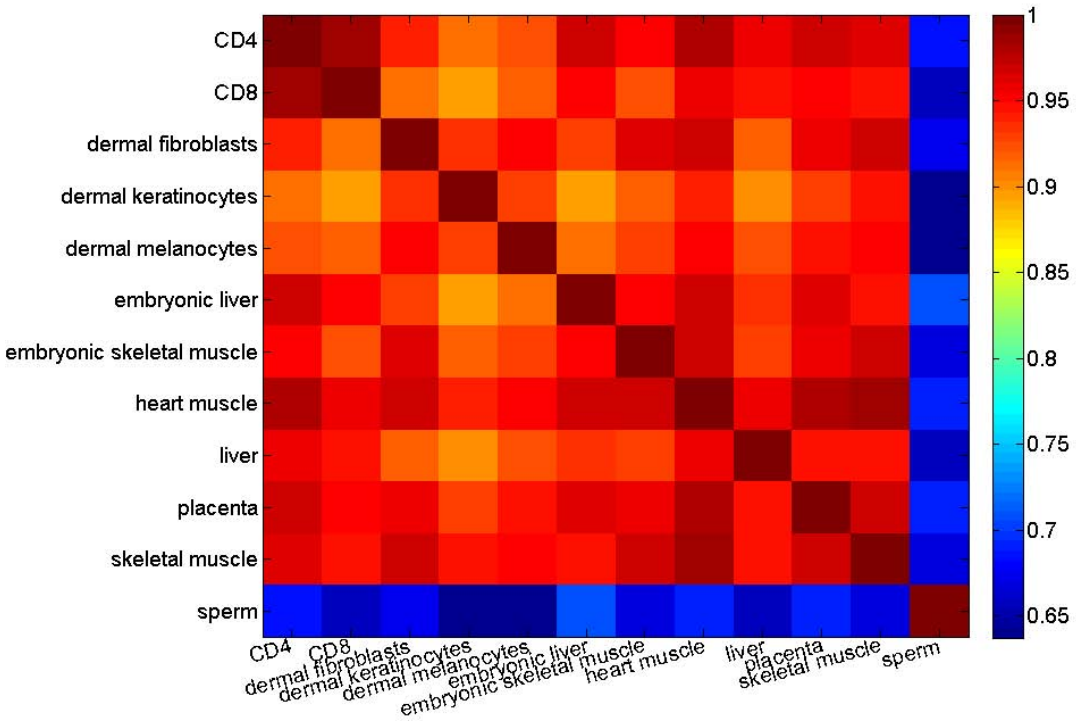

Figure 4: Correlation coefficients of the CGI methylation levels across different tissues and cell types.

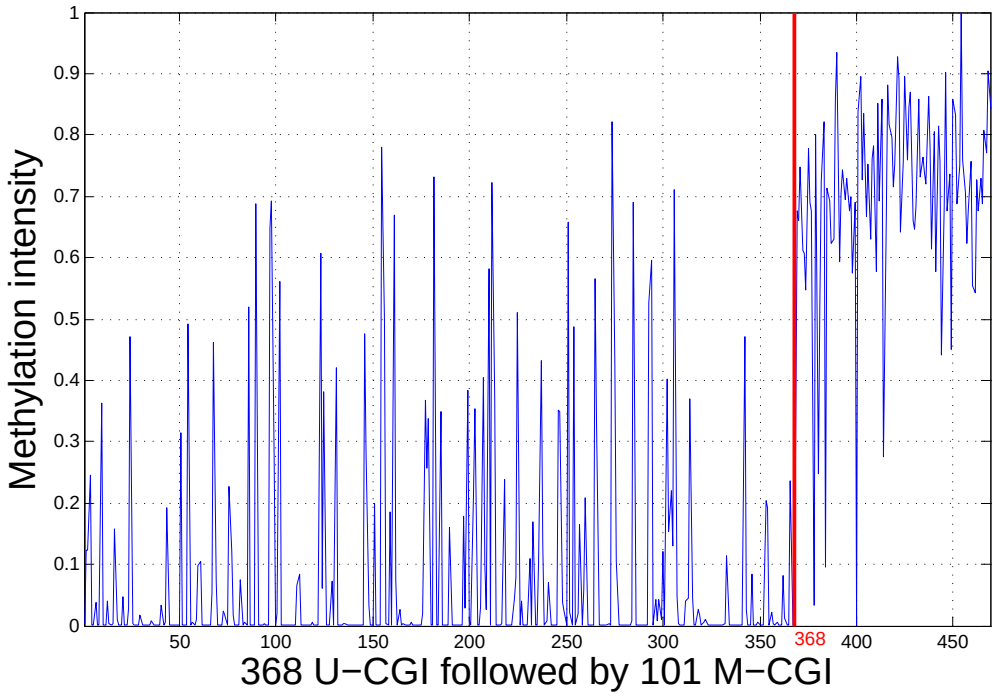

Figure 5: Methylation intensity generated by the regression analysis for the 368 unmethylated CGIs and 101 methylated CGIs.

6.3 Aberrant Methylation Prediction

We first selected 78 CGIs that are differentially methylated and whose differential methylation are related to some cancer types, and 783 CGIs that are constantly unmethylated across various tissues, cells and phenotypes based on the data from MethCancerDB [64], mPod [47][48], and our computational model [93]. Out of the 640 features regarding (1) CGI specific attributes, (2) sequence composition of the CGI, (3) structure of the CGI, and (4) distribution of TFBSs and conserved elements in or near the CGI, we obtained through statistical tests 88 features having different distribution between differentially methylated and constantly unmethylated CGIs. These 88 features include one CGI specific attribute, 35 DNA composition features, seven DNA structure features, 43 TFBS-related features, and two evolutionarily conserved element-related features. Between the previous 342 features that are statistically different between methylated and unmethylated CGIs and these 88

features that are statistically different between differentially methylated and constantly unmethylated CGIs, only 30 are in common. The first 33 principal components that can retain 99.99% variance in the original feature space were selected through PCA as shown in Table 11. Our results further demonstrated that the support vector machine-based classifier with the reduced dimensional feature space rendered from PCA could achieve a ∼70% specificity, ∼73% sensitivity, and ∼71% accuracy in distinguishing the differentially methylated from the constantly unmethylated CGIs.

Table 11: Number of genes, statistically significant features and PCs for different cases of aberrant methylation prediction. Performance was measured using specificity, sensitivity and accuracy.

Case	**Statistics**			**Performance**		
	# genes	**# features**	**# PCs**	**SP**	**SE**	**ACC**
any cancer	78	88	33	0.70	0.73	0.71
colon	177	75	40	0.99	0.92	0.92
prostate	122	45	21	0.98	0.93	0.93

We then investigated 177 genes aberrantly methylated in colon cancer, treating the housekeeping genes as the control. Through statistical tests, we identified 75 features having different distribution between aberrantly methylated and constantly unmethylated CGIs. These 75 features include one CGI specific attribute, 58 DNA composition features, eight DNA structure features, six TFBS-related features, and two evolutionarily conserved element-related features. By using the first 40 principal components that retains 99.99% of the variance, our support vector machine-based classifier can reach ∼99% specificity, ∼92% sensitivity, and ∼92% accuracy in distinguishing the aberrantly methylated in colon cancer from the constantly unmethylated CGIs. Fig. 6 shows the histogram of the predicted scores of all promoter CGIs using our computational predictive model for the potential of aberrant methylation in colon cancer. As can be seen from Fig. 6, the predicted scores for the colon cancer genes are concentrated in the lower range and the scores for the housekeeping genes are concentrated in the higher range. Specifically, ∼83% of the colon cancer related genes lie below the score 0.3, and ∼91% of the housekeeping gene CGIs lie above the score 0.7. This prompts us to prioritize those CGIs with low predicted scores for further experimental

validation as the aberrant methylation targets in colon cancer.

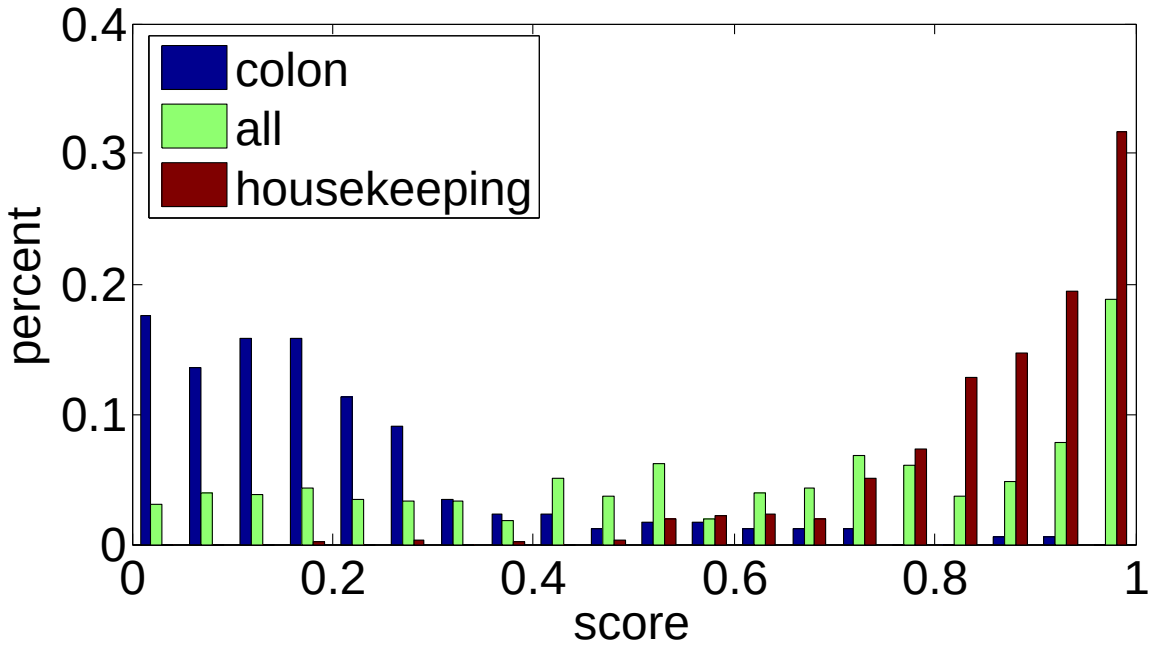

Figure 6: Histogram of the predicted scores of all promoter CGIs for the potential of aberrant methylation in colon cancer.

Likewise, for the 122 genes aberrantly methylated in prostate cancer, we identified 45 features showing statistical difference between aberrantly methylated and constantly unmethylated CGIs. The identified features include two CGI specific attributes, 31 DNA composition features, eight DNA structure features, two TFBS-related features, and two evolutionarily conserved element-related features. By using the PCA-based model constructed from the first 21 principal components, we can achieve ∼98% specificity, ∼93% sensitivity, and ∼93% accuracy in the classification. Similar observations can be made from Fig. 7 for the prostate cancer as for the colon cancer. The predicted scores for the prostate cancer genes and the housekeeping genes are concentrated in the two extreme ends. Specifically, ∼85% of the prostate cancer related genes lie below the score 0.3, and ∼90% of the housekeeping gene CGIs lie above the score 0.7. We can thus prioritize those CGIs with low predicted scores for further biological experiment validation as the aberrant methylation targets in prostate cancer.

Since some cancer-related aberrantly methylated genes are shared by these two cancer

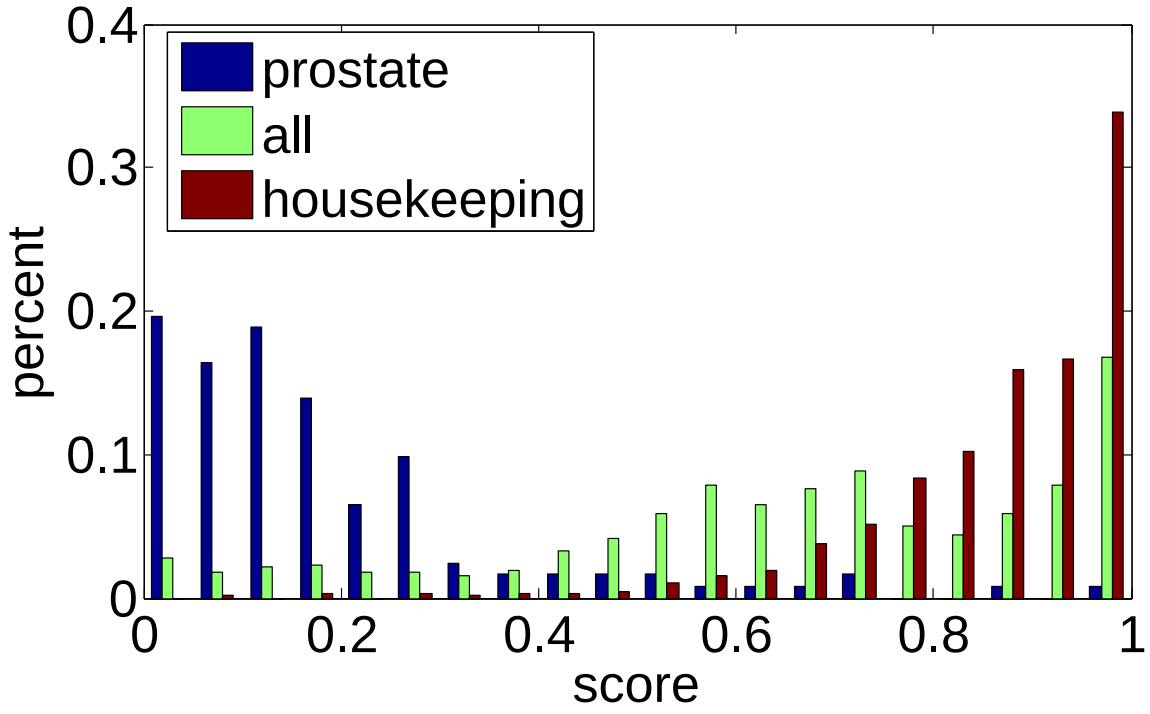

Figure 7: Histogram of the predicted scores of all promoter CGIs for the potential of aberrant methylation in prostate cancer.

types, we further investigated the performance of our predictive models on the common genes and non-common genes of the colon and prostate cancer. Table 12 summarized the number of common and non-common genes between the two cancer types, as well as the accuracy of the predictive models on all colon or prostate cancer related genes, the common genes, and non-common genes. Observe from Table 12 that the model built from the colon cancer genes can achieve better accuracy on the colon cancer genes than on the prostate cancer genes. Particularly, an accuracy of ∼95% can be achieved for the colon cancer specific genes while the accuracy drops to ∼71% on the prostate-cancer specific genes. Similar observations can be made from the accuracy of the model constructed from the the prostate cancer related genes. An accuracy of ∼93% can be achieved for the prostate cancer specific genes while the accuracy drops to ∼75% on the colon cancer specific genes. This indicates that our aberrantly methylation predictive model trained on a particular cancer type is more specific to detect the aberrant methylation in that particular cancer type, and the prediction accuracy would decrease when applying the model to a different

cancer type. These observations suggest that different mechanisms exist in which DNA methylation contributes to different cancer formation.

Table 12: Accuracy of the aberrant methylation predictive models on all colon or prostate cancer related genes, the common genes shared by the two cancers types, and genes specific to a cancer type. The number in parentheses denotes the number of CGIs within the category.

Model	**colon**			**prostate**		
	all (177)	**specific (83)**	**common (94)**	**all (122)**	**specific (28)**	**common (94)**
colon	0.9209	0.9518	0.8936	0.8525	0.7143	0.8936
prostate	0.8427	0.7470	0.9255	0.9262	0.9286	0.9255

CHAPTER VII

CONCLUSION AND FUTURE WORK

7.1 Conclusion

DNA methylation is a type of epigenetic modification which involves the addition of a methyl group to DNA via DNMT. The establishment of DNA methylation pattern is a crucial part of cell differentiation and organ development, suppression of viral genes and deleterious elements, and carcinogenesis. Computational predictions of DNA methylation profiles serve multiple purposes. First, accurate predictions can contribute valuable information for speeding up genome-wide DNA methylation profiling so that experimental resources can be focused on a few selected while computational procedures are applied to the bulk of the genome. Second, computational predictions can extract functional features and construct useful models of DNA methylation based on existing data, and can therefore be used as an initial step toward quantitative identification of critical factors or pathways controlling DNA methylation patterns. Third, computational prediction of DNA methylation can provide benchmark data to calibrate DNA methylation profiling equipment and to consolidate profiling results from different techniques or equipments.

We studied the computational analysis of the DNA methylation patterns in human genome. We incorporated multiple information resources, including (1) HEP and mPod that specify DNA methylation information in normal tissues, (2) MethCancerDB that specify aberrant methylation information in cancerous conditions, (3) CGC, AGCOH, CancerGenes and OMIM for functional annotation, and (4) housekeeping genes from literature, for the analysis. We designed and implemented various statistical tests, PCA, and model development methods, and achieved predictive models with high performance.

We have identified various features that are associated with the methylation status of the CGIs in human genome. We constructed support vector machine-based classifier models to discriminate between CGIs that are prone to methylation from those that are resistant to

methylation based on the identified features. In comparison with the existing methods, our models can achieve higher performance with an accuracy of 93-94%, specificity of 94%, and sensitivity of 92-93% for predicting CGI methylation status in normal CD4 lymphocytes from the HEP data set. We showed that our models, constructed from CD4 lymphocytes, can be applied to predict the methylation status of some other normal tissues and cell types. To profile the methylation intensity of CGIs in normal tissues, we also built support vector machine-based regression models, which generate continuous intensity values. We showed that the intensity values generated by the regression models correlated well with the intensities values based on the HEP data set (correlation coefficient $\rho = 0.883$, p-value$<4\times10^{-122}$).

We also investigated patterns indicative of methylation variation in normal tissues versus cancerous tissues. We performed the analysis of aberrant methylation under two different settings: aberrant methylation in any type of cancer, and aberrant methylation in a specific cancer type. For the latter setting, we considered the colon and prostate cancer. We correlated various features with cancer related aberrant methylation of CGI in these two settings. More specifically, We identified 88 statistically significant features between aberrantly methylated and consistently unmethylated CGIs in cancerous conditions. We also found 75 and 45 features showing differential patterns between aberrantly methylated and consistently unmethylated CGIs for colon and prostate cancer, respectively. Furthermore, based on these differential features, we built predictive models to detect such aberrantly methylated CGIs in cancer via various modified cross-validation tests and generalization tests. Experimental results showed that our predictive models can achieve high accuracy (92-93%), specificity (98-99%) as well as sensitivity (92-93%) for both the colon and prostate cancer. We also used our predictive models to all promoter CGIs in human genome to prioritize potentially aberrantly methylated genes in cancer.

The major contribution of our study lies in three aspects. First, we detected various genetic and epigenetic features that are associated with the methylation status and cancer related aberrant methylation of the CGIs in human genome. Such features can serve as the foundation for exploring the exact mechanisms of DNA methylation in normal organismal

development and cancerogenesis. Second, our DNA methylation predictive model can serve as a fast and effective way to explore genome-wide CGI methylation profiles in normal tissues. Third, our predictive model for cancer related aberrant methylation can serve as an initial step to prioritize and detect novel aberrantly methylated genes in cancerous conditions.

7.2 Future Work

With the advance of experimental technologies for genome-wide DNA methylation profiling and the expected bombardment of DNA methylation data, more challenges will be posed on the subsequent computational epigenomics analysis to extract and abstract useful and consistent information across these large amounts of data. Advanced techniques to effectively process and integrate data of different resources and of different scales are in great need to improve existing models or to establish novel models for DNA methylation and related epigeneitc and genetic phenomena.

Though it is known that DNA methylation is heavily involved in the normal development and differentiation, as well as in the onset and progression of diseases, the exact mechanisms are yet to be discovered. It will certainly help to accelerate biomedical investigations if we can, through computational predictions, comparative analyses, and evolutionary studies, identify those DNA regions whose methylation variation patterns are correlated with, indicative of, and underlying of the variations in gene expressions, histone modifications and chromatin structures that are related to normal development, cell differentiation, genome imprinting, X-chromosome inactivation, and phenotypic changes, respectively.

An improved understanding of the DNA methylation mechanisms, especially about the DNA methylation variations with respect to other genetic or epigenetic modifications, will help move towards a prospective DNA methylation-based pharmacology. Further challenges may include modulating cancer growth and metastasis by targeting different proteins of the DNA methylation machinery to achieve a balance of anticancer therapy with positive outcome and reduced side effects [122][123].

CHAPTER VIII

APPENDIX

8.1 Abbreviated Terms

The abbreviations used in this dissertation are summarized in Table 13.

Table 13: List for the abbreviations used in the thesis and their corresponding full names.

Abbreviation	**Full Name**
ACC	Accuracy
bp	Base Pair
CGC	Cancer Gene Census
CGI	CpG Island
CpG	Cytosine-phosphate-Guanine
DMR	Differentially Methylated Region
DNA	Deoxyribonucleic acid
DNMT	DNA methyltransferase
GO	Gene Ontology
HEP	Human Epigenome Project
LOOCV	Leave-One-Out Cross-Validation
Methyl-MAPS	Methylation Mapping Analysis by Paired-end Sequencing
MDRE	Methylation Dependent Restriction Enzyme
MeDIP	Methylated DNA Immunoprecipitation
mPod	Methylation Profiles of DNA
MSRE	Methylation Sensitive Restriction Enzyme
NCBI	National Center for Biotechnology Information
OMIM	Online Mendelian Inheritance in Man
PCA	Principal Component Analysis
REBASE	Restriction Enzyme Database
ROC	Receiver Operating Characteristic
ROI	Region of Interest
SE	Sensitivity
SP	Specificity
SNP	Single-nucleotide Polymorphism
TFBS	Transcription Factor Binding Site
TSG	Tumor Suppressor Gene
UCSC	University of California, Santa Cruz

8.2 Enrichment Analysis

We examined whether a CGI's nearby genes are involved in any cancer-related biological processes. Altogether, we have identified 11 biological processes that are enriched by TSGs, and 30 biological processes that are enriched by oncogenes. Four biological processes are enriched by both TSGs and oncogenes. These cancer-related biological processes are listed as follows in Tables 14 and 15, with the four commonly enriched biological processes highlight in italics.

Table 14: Oncogene enriched biological processes with their GO identifiers and enrichment factors.

Gene Ontology	Name	Enrich Factor
GO:0006355	regulation of transcription, DNA-dependent	1.8302
GO:0006350	transcription	1.8709
GO:0045449	regulation of transcription	2.411
GO:0007275	multicellular organismal development	1.3131
GO:0006468	protein amino acid phosphorylation	1.7815
GO:0006366	transcription from RNA polymerase II promoter	3.2426
GO:0007049	cell cycle	1.1148
GO:0007169	transmembrane receptor protein tyrosine kinase signaling	6.1067
GO:0008284	positive regulation of cell proliferation	2.7833
GO:0008283	cell proliferation	1.4073
GO:0030154	cell differentiation	1.0006
GO:0045944	+ reg transcription from RNA polymerase II promoter	3.1473
GO:0006916	anti-apoptosis	2.3605
GO:0007166	cell surface receptor linked signaling pathway	1.7167
GO:0009887	organ morphogenesis	3.2937
GO:0045941	positive regulation of transcription	4.6435
GO:0001501	skeletal system development	2.6503
GO:0007242	intracellular signaling pathway	1.0491
GO:0042981	regulation of apoptosis	2.8612
GO:0051301	cell division	1.5936
GO:0000122	- reg transcription from RNA polymerase II promoter	2.1184
GO:0006461	protein complex assembly	1.8513
GO:0006897	endocytosis	2.2481
GO:0006974	response to DNA damage stimulus	1.2262
GO:0045893	positive regulation of transcription, DNA-dependent	2.8185
GO:0009653	anatomical structure morphogenesis	1.8513
GO:0016481	negative regulation of transcription	2.2164
GO:0016568	chromatin modification	1.3926
GO:0030097	hemopoiesis	5.6201
GO:0043123	positive regulation of I-kappaB kinase/NF-kappaB cascade	2.0437

Table 15: TSG enriched biological processes with their GO identifiers and enrichment factors.

Gene Ontology	**Name**	**Enrich Factor**
GO:0007049	cell cycle	5.1818
GO:0045786	negative regulation of cell cycle	17.8968
GO:0006281	DNA repair	8.9897
GO:0006974	response to DNA damage stimulus	9.9741
GO:0008285	negative regulation of cell proliferation	3.612
GO:0006289	nucleotide-excision repair	25.6002
GO:0006461	protein complex assembly	3.6879
GO:0007050	cell cycle arrest	6.5502
GO:0006298	mismatch repair	19.2483
GO:0006917	induction of apoptosis	3.4179
GO:0045944	+ reg transcription from RNA polymerase II promoter	3.3247

REFERENCES

[1] A.P. Bird. CpG-rich islands and the function of DNA methylation. *Nature*, 321:209–213, 1986.

[2] E.L. Kinnally, C. Feinberg, D. Kim, K. Ferguson, R. Leibel, J.D. Coplan, and J. John Mann. DNA methylation as a risk factor in the effects of early life stress. *Brain Behav Immun.*, 25:1548–53, 2011.

[3] S.C. Galvan, M. Martinez-Salazar, V.M. Galvan, R. Mendez, G.T. Diaz-Contreras, M. Alvarado-Hermida, R. Alcantara-Silva, and A. Garcia-Carranca. Analysis of CpG methylation sites and CGI among human papillomavirus DNA genomes. *BMC Genomics*, 12:580, 2011.

[4] M. Ehrlich, M.A. Gama-Sosa, L. Huang, R.M. Midgett, K.C. Kuo, R.A. McCune, and C. Gehrke. Amount and distribution of 5-methylcytosine in human DNA from different types of tissues or cells. *Nucleic Acids Research*, 10:2709–2721, 1982.

[5] M. Gardiner-Garden and M. Frommer. CpG islands in vertebrate genomes. *Journal of molecular biology*, 196:261–282, 1987.

[6] C. Bock, M. Paulsen, S. Tierling, T. Mikeska, T. Lengauer, and J. Walter. CpG island methylation in human lymphocytes is highly correlated with DNA sequence, repeats, and predicted DNA structure. *PLoS Genetics*, 2:e26, 2006.

[7] F. Sleutels, R. Zwart, and D.P. Barlow. The noncoding air RNA is required for silencing autosomal imprinted genes. *Nature*, 415:820, 2002.

[8] L. Han, B. Su, W. Li, and Z. Zhao. CpG island density and its correlations with genomic features in mammalian genomes. *Genome Biology*, 9:R79, 2008.

[9] F.A. Feltus, E.K. Lee, J.F. Costello, C. Plass, and P.M. Vertino. Predicting aberrant CpG island methylation. *Proceedings of the National Academy of Sciences USA*, 100:12253–12258, 2003.

[10] F. Antequera and A. Bird. Number of CpG islands and genes in human and mouse. *Proceedings of the National Academy of Sciences USA*, 90:11995–9, 1993.

[11] P.P. Lee, D.R. Fitzpatrick, C. Beard, H.K. Jessup, S. Lehar, K.W. Makar, M. Pérez-Melgosa, M.T. Sweetser, M.S. Schlissel, S. Nguyen, S.R. Cherry, J.H. Tsai, S.M. Tucker, W.M. Weaver, A. Kelso, R. Jaenisch, and C.B. Wilson. A critical role for Dnmt1 and DNA methylation in T cell development, function, and survival. *Immunity.*, 15:763–74, 2001.

[12] R.D. Hawkins, G.C. Hon, L.K. Lee, Q. Ngo, R. Lister, M. Pelizzola, L.E. Edsall, S. Kuan, Y. Luu, S. Klugman, J. Antosiewicz-Bourget, Z. Ye, C. Espinoza, S. Agarwahl, L. Shen, V. Ruotti, W. Wang, R. Stewart, J.A. Thomson, J.R. Ecker, and B. Ren. Distinct epigenomic landscapes of pluripotent and lineage-committed human cells. *Cell Stem Cell.*, 6:479–91, 2010.

[13] N.S. Christophersen and K. Helin. Epigenetic control of embryonic stem cell fate. *J Exp Med.*, 207:2287–2295, 2011.

[14] R.L. Poole, D.J. Leith, L.E. Docherty, M.E. Shmela, C. Gicquel, M. Splitt, I.K. Temple, and D.J.G. Mackay. BeckwithcWiedemann syndrome caused by maternally inherited mutation of an OCT-binding motif in the IGF2/H19-imprinting control region, ICR1. *European Journal of Human Genetics*, doi:10.1038/ejhg.2011.166, 2011.

[15] A.M. Cotton, L. Lam, J.G. Affleck, I.M. Wilson, M.S. Peñaherrera, D.E. McFadden, M.S. Kobor, W.L. Lam, W.P. Robinson, and C.J. Brown. Chromosome-wide DNA methylation analysis predicts human tissue-specific X inactivation. *Hum Genet.*, 130:187–201, 2011.

[16] K.L. Novik, I. Nimmrich, B. Genc, S. Maier, C. Piepenbrock, A. Olek, and S. Beck. Epigenomics: Genome-wide study of methylation phenomena. *Curr. Issues Mol. Biol.*, 4:111–128, 2002.

[17] S.W. Jiang, J. Li, K. Podratz, and S. Dowdy. Application of DNA methylation biomarkers for endometrial cancer management. *Expert Rev Mol Diagn.*, 8:607–16, 2008.

[18] V. Greger, E. Passarge, W. Höpping, E. Messmer, and B. Horsthemke. Epigenetic changes may contribute to the formation and spontaneous regression of retinoblastoma. *Hum Genet.*, 83:155–8, 1989.

[19] J. Reibenwein, D. Pils, P. Horak, B. Tomicek, G. Goldner, N. Worel, K. Elandt, and M. Krainer. Promoter hypermethylation of GSTP1, AR, and 14-3-3sigma in serum of prostate cancer patients and its clinical relevance. *Prostate*, 67:427–32, 2007.

[20] S. Nambu, K. Inoue, and H. Saski. Site-specific hypomethylation of the c-myc oncogene in human hepatocellular carcinoma. *Jpn J Cancer Res.*, 78:695–704, 1987.

[21] C. Shao, W. Sun, M. Tan, C.A. Glazer, S. Bhan, X. Zhong, C. Fakhry, R. Sharma, W.H. Westra, M.O. Hoque, C.A. Moskaluk, D. Sidransky, J.A. Califano, and P.K. Ha. Integrated, genome-wide screening for hypomethylated oncogenes in salivary gland adenoid cystic carcinoma. *Clin Cancer Res.*, 17:4320–30, 2011.

[22] J. Tost. DNA methylation: An introduction to the biology and the disease-associated changes of a promising biomarker. *Mol Biotechnol*, 44:71–81, 2010.

[23] L. Hartnett and L.J. Egan. Inflammation, DNA methylation and colitis-associated cancer. *Carcinogenesis*, online:doi: 10.1093/carcin/bgs006, 2012.

[24] S. Beck and V.K. Rakyan. The methylome: approaches for global dna methylation profiling. *Trends Genet.*, 24:231–7, 2008.

[25] H. Zheng, S-W. Jiang, and H. Wu. A review on the techniques for characterizing and predicting human genomic DNA methylation. *Current Bioinformatics*, 2012.

[26] I.M. Carr, E.M.A. Valleley, S.F. Cordery, A.F. Markham, and D.T. Bonthron. Sequence analysis and editing for bisulphite genomic sequencing projects. *Nucl. Acids Res.*, 35:e79, 2007.

[27] J. Reinders, C.D. Vivier, G. Theiler, D. Chollet, P. Descombes, and J. Paszkowski. Genome-wide, high-resolution DNA methylation profiling using bisulfite-mediated cytosine conversion. *Genome Res.*, 18:469–476, 2008.

[28] A. Meissner, T.S. Mikkelsen, H. Gu, M. Wernig, J. Hanna, A. Sivachenko, X. Zhang, B.E. Bernstein, C. Nusbaum, D.B. Jaffe, A. Gnirke, R. Jaenisch, and E.S. Lander. Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature*, 454:766–70, 2008.

[29] R. Lister, M. Pelizzola, R.H. Dowen, R.D. Hawkins, G. Hon, J. Tonti-Filippini, J.R. Nery, L. Lee, Z. Ye, Q.M. Ngo, L. Edsall, J. Antosiewicz-Bourget, R. Stewart, V. Ruotti, A.H. Millar, J.A. Thomson, B. Ren, and J.R. Ecker. Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature*, 462:315–22, 2009.

[30] Y. Korshunova, R.K. Maloney, N. Lakey, R.W. Citek, B. Bacher, A. Budiman, J.M. Ordway, W.R. McCombie, J. Leon, J.A. Jeddeloh, and J.D. McPherson. Massively parallel bisulphite pyrosequencing reveals the molecular complexity of breast cancer-associated cytosine-methylation patterns obtained from tissue and serum DNA. *Genome Res*, 18:19–29, 2008.

[31] C.A. Bormann Chung, V.L. Boyd, K.J. McKernan, Y. Fu, C. Monighetti, H.E. Peckham, and M. Barker. Whole methylome analysis by ultra-deep sequencing using two-base encoding. *PLoS One.*, 5:e9320, 2010.

[32] R.P. Darst, C.E. Pardo, K.D. Ai, L amd Brown, and M.P. Kladde. Bisulfite sequencing of DNA. *Current Protocols in Molecular Biology*, 7.9:1–17, 2010.

[33] R. Gupta, A. Nagarajan, and N. Wajapeyee. Advances in genome-wide DNA methylation analysis. *Biotechniques*, 49:iii–xi, 2010.

[34] `http://www.neb.com/nebecomm/tech_reference/restriction_enzymes/dam_dcm_cpg_methylation.asp`.

[35] F.J. Stewart, D. Panne, T.A. Bickle, and E.A. Raleigh. Methyl-specific DNA binding by McrBC, a modification-dependent restriction enzyme. *J Mol Biol.*, 12:611–22, 2000.

[36] R.J. Roberts, T. Vincze, J. Posfai, and D. Macelis. REBASE–a database for DNA restriction and modification: enzymes, genes and genomes. *Nucl. Acids Res.*, 38:D234–D236, 2010.

[37] M.Q. Zhang and A.D. Smith. Challenges in understanding genome-wide DNA methylation. *Journal of Computer Science and Technology*, 25:26–34, 2010.

[38] E.K. Ng, C.P. Leung, V.Y. Shin, C.L. Wong, E.S. Ma, H.C. Jin, K.M. Chu, and A. Kwong. Quantitative analysis and diagnostic significance of methylated SLC19A3 DNA in the plasma of breast and gastric cancer patients. *PLoS One.*, 6:e22233, 2011.

[39] K. Hashimoto, S. Kokubun, E. Itoi, and H.I. Roach. Improved quantification of DNA methylation using methylation-sensitive restriction enzymes and real-time PCR. *Epigenetics*, 2:86–91, 2007.

[40] D.J. Ciavatta, J. Yang, G.A. Preston, A.K. Badhwar, H. Xiao, P. Hewins, C.M. Nester, W.F. 3rd Pendergraft, T.R. Magnuson, J.C. Jennette, and R.J. Falk. Epigenetic basis for aberrant upregulation of autoantigen genes in humans with ANCA vasculitis. *J Clin Invest.*, 120:3209–19, 2010.

[41] Z. Lippman, A.V. Gendrel, V. Colot, and R. Martienssen. Profiling DNA methylation patterns using genomic tiling microarrays. *Nature Methods*, 2:219–224, 2005.

[42] J.R. Edwards, A.H. O'Donnell, R.A. Rollins, H.E. Peckham, C. Lee, M.H. Milekic, B. Chanrion, Y. Fu, T. Su, H. Hibshoosh, J.A. Gingrich, F. Haghighi, R. Nutter, and T.H. Bestor. Chromatin and sequence features that define the fine and gross structure of genomics methylation patterns. *Genome Res.*, 20:972–80, 2010.

[43] M. Weber, J.J. Davies, D. Wittig, E.J. Oakeley, M. Haase, W.L. Lam, and D. Schubeler. Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells. *Nature Genetics*, 37:853–862, 2005.

[44] F. Mohn, M. Weber, D. Schübeler, and T.C. Roloff. Methylated DNA immunoprecipitation (MeDIP). *Angewandte Chemie*, 50:6460C6468, 2011.

[45] K.L. Thu, E.A. Vucic, J.Y. Kennett, C. Heryet, C.J. Brown, W.L. Lam, and I.M. Wilson. Methylated DNA immunoprecipitation. *J Vis Exp.*, 23:935, 2009.

[46] M. Weber, I. Hellmann, M. Stadler, L. Ramos, S. Paabo, M. Rebhan, and D. Schubeler. Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome. *Nature Genetics*, 39:457–466, 2007.

[47] V.K. Rakyan, T.A. Down, N.P. Thorne, P. Flicek, E. Kulesha, S. Graf, E.M. Tomazou, L. Backdahl, N. Johnson, M. Herberth, K.L. Howe, D.K. Jackson, M.M. Miretti, H. Fiegler, J.C. Marioni, E. Birney, T.J.P. Hubbard, N.P. Carter, S. Tavare, and S. Beck. An integrated resource for genome-wide identification and analysis of human tissue-specific differentially methylated regions (tDMRs). *GENOME RES*, 18:1518–1529, 2008.

[48] T.A. Down, V.K. Rakyan, D.J. Turner, P. Flicek, H. Li, E. Kulesha, S. Gräf, N. Johnson, J. Herrero, E.M. Tomazou, N.P. Thorne, L. Bäckdahl, M. Herberth, K.L. Howe, D.K. Jackson, M.M. Miretti, J.C. Marioni, E. Birney, T.J. Hubbard, R. Durbin, S. Tavaré, and S. Beck. A bayesian deconvolution strategy for immunoprecipitation-based DNA methylome analysis. *Nat Biotechnol.*, 26:779–85, 2008.

[49] K.R. Pomraning, K.M. Smith, and M. Freitag. Genome-wide high throughput analysis of DNA methylation in eukaryotes. *Methods*, 47:142–50, 2009.

[50] Y.M.D. Lo, R.W.K. Chiu, and K.C.A Chan. *Clinical applications of PCR.* Humana Press, 2nd edition, 2006.

[51] B. Weinhold. Epigenetics: The science of change. *Environ Health Perspect*, 114:A160–A167, 2006.

[52] `http://www.epigenome.org/index.php?page=project`.

[53] V.K. Rakyan, T. Hildmann, K.L. Novik, J. Lewin, J. Tost, A.V. Cox, T.D. Andrews, K.L. Howe, T. Otto, A. Olek, J. Fischer, I.G. Gut, K. Berlin, and S. Beck. DNA methylation profiling of the human major histocompatibility complex: A pilot study for the human epigenome project. *PLoS Biol*, 2:e405, 2004.

[54] F. Eckhardt, J. Lewin, R. Cortese, V.K. Rakyan, J. Attwood, M. Burger, J. Burton, T.V. Cox, R. Davies, T.A. Down, C. Haefliger, R. Horton, K. Howe, D.K. Jackson, J. Kunde, C. Koenig, J. Liddle, D. Niblett, T. Otto, R. Pettett, S. Seemann, C. Thompson, T. West, J. Rogers, A. Olek, K. Berlin, and S. Beck. DNA methylation profiling of human chromosomes 6, 20 and 22. *Nature Genetics*, 38:1378–1385, 2006.

[55] J.T. Bell, A.A. Pai, J.K. Pickrell, D.J. Gaffney, R. Pique-Regi, J.F. Degner, Y. Gilad, and J.K. Pritchard. DNA methylation patterns associate with genetic and gene expression variation in HapMap cell lines. *Genome Biol.*, 12:R10, 2011.

[56] Z.A. Kaminsky, T. Tang, S.C. Wang, C. Ptak, G.H. Oh, A.H. Wong, L.A. Feldcamp, C. Virtanen, J. Halfvarson, C. Tysk, A.F. McRae, P.M. Visscher, G.W. Montgomery, I.I. Gottesman, N.G. Martin, and A. Petronis. DNA methylation profiles in monozygotic and dizygotic twins. *Nat Genet.*, 41:240–5, 2009.

[57] A. Schumacher, P. Kapranov, Z. Kaminsky, J. Flanagan, A. Assadzadeh, P. Yau, C. Virtanen, N. Winegarden, J. Cheng, T. Gingeras, and A. Petronis. Microarray-based DNA methylation profiling: technology and applications. *Nucleic Acids Res.*, 34:528–542, 2006.

[58] J. Flanagan, V. Popendikyte, N. Pozdniakovaite, M. Sobolev, A. Assadzadeh, A. Schumacher, M. Zangeneh, L. Lau, C. Virtanen, S-C. Wang, and A. Petronis. Intra- and inter- individual epigenetic variation in human germ cells. *American Journal of Human Genetics*, 79:67–84, 2006.

[59] `http://www.methylogix.com/genetics/database.shtml.htm`.

[60] D. Ratel, J. Ravanat, F. Berger, and D. Wion. N6-methyladenine: the other methylated base of DNA. *Bioessays*, 28:309–315, 2006.

[61] S. Yagi, K. Hirabayashi, S. Sato, W. Li, Y. Takahashi, T. Hirakawa, G. Wu, N. Hattori, N. Hattori, J. Ohgane, S. Tanaka, X.S. Liu, and K. Shiota. DNA methylation profile of tissue-dependent and differentially methylated regions (T-DMRs) in mouse promoter regions demonstrating tissue-specific gene expression. *Genome Res*, 18:1969–78, 2008.

[62] F. Pattyn, J. Hoebeeck, P. Robbrecht, E. Michels, A. De Paepe, G. Bottu, D. Coornaert, R. Herzog, F. Speleman, and J. Vandesompele. methblast and methprimerdb: web-tools for PCR based methylation analysis. *BMC Bioinformatics*, 7:496, 2006.

[63] V. Negre and C. Grunau. The MethDB DAS server: adding an epigenetic information layer to the human genome. *Epigenetics.*, 1:101–5, 2006.

[64] M. Lauss, I. Visne, A. Weinhaeusel, K. Vierlinger, C. Noehammer, and A. Kriegner. MethcancerDB - aberrant DNA methylation in human cancer. *British Journal of Cancer*, 98:816–817, 2008.

[65] M. Ongenaert, L. Van Neste, T. De Meyer, G. Menschaert, S. Bekaert, and W. Van Criekinge. Pubmeth: a cancer methylation database combining text-mining and expert annotation. *Nucleic Acids Res.*, 36(Database issue):D842–6, 2008.

[66] Y.C. Fang, H.C. Huang, and H.F. Juan. MeInfoText: associated gene methylation and cancer information from text mining. *BMC Bioinformatics*, 9:22, 2008.

[67] X. He, S. Chang, J. Zhang, Q. Zhao, H. Xiang, K. Kusonmano, L. Yang, Z.S. Sun, H. Yang, and J. Wang. MethyCancer: the database of human DNA methylation and cancer. *Nucleic Acids Res.*, 36(Database issue):D836–41, 2008.

[68] C. Grunau, E. Renault, A. Rosenthal, and G. Roizes. Methdba public database for DNA methylation data. *Nucleic Acids Res.*, 29:270–274, 2001.

[69] P.W. Laird. The power and the promise of DNA methylation markers. *Nature Reviews Cancer*, 3:253–266, 2003.

[70] A. Barat and H. Ruskin. A manually curated novel knowledge management system for genetic and epigenetic molecular determinants of colon cancer. *The Open Colorectal Cancer Journal*, 3:36–46, 2010.

[71] A.P. Bird. DNA methylation patterns and epigenetic memory. *Genes & Development*, 16:6–21, 2002.

[72] Y. Yamada, H. Watanabe, F. Miura, H. Soejima, M. Uchiyama, T. Iwasaka, T. Mukai, Y. Sakaki, and T. Ito. A comprehensive analysis of allelic methylation status of CpG islands on human chromosome 21q. *Genome Research*, 14:247–266, 2004.

[73] L. Vrba, J.C. Garbe, M.R. Stampfer, and B.W. Futscher. Epigenetic regulation of normal human mammary cell type-specific miRNAs. *Genome Res.*, 21:2026–37, 2011.

[74] F. Fang, S. Fan, X. Zhang, and M.Q. Zhang. Predicting methylation status of CpG islands in the human brain. *Bioinformatics*, 22:2204–2209, 2006.

[75] M. Bhasin, H. Zhang, E.L. Reinherz, and P.A. Reche. Prediction of methylated CpGs in DNA sequences using a support vector machine. *FEBS Lett*, 579:4302–8, 2005.

[76] L. Lu, K. Lin, Z. Qian, H. Li, Y. Cai, and Y. Li. Predicting DNA methylation status using word composition. *J. Biomedical Science and Engineering*, 3:672–676, 2010.

[77] C. Bock, J. Walter, M. Paulsen, and T. Lengauer. CpG island mapping by epigenome prediction. *PLoS Computational Biology*, 3:e110, 2007.

[78] I. Ali and H. Seker. Detailed methylation prediction of CpG islands on human chromosome 21. *10th WSEAS International Conference on Mathematics and Computers In Biology and Chemistry*, pages 147–152, 2009.

[79] S. Fan, M.Q. Zhang, and X. Zhang. Histone methylation marks play important roles in predicting the methylation status of CpG islands. *Biochemical and Biophysical Research Communications*, 374:559–564, 2008.

[80] C. Previti, O. Harari, I. Zwir, and C. del Val. Profile analysis and prediction of tissue-specific CpG island methylation classes. *BMC Bioinformatics*, 10:116, 2009.

[81] J. Lv, J. Su, F. Wang, Y. Qi, H. Liu, and Y. Zhang. Detecting novel hypermethylated genes in breast cancer benefiting from feature selection. *Comput Biol Med.*, 40:159–67, 2010.

[82] A. Siepel, G. Bejerano, J.S. Pedersen, A. Hinrichs, M. Hou, K. Rosenbloom, H. Clawson, J. Spieth, L.W. Hillier, S. Richards, G.M. Weinstock, R.K. Wilson, R.A. Gibbs, W.J. Kent, W. Miller, and D. Haussler. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Research*, 15:1034–1050, 2005.

[83] B.E. Bernstein, T.S. Mikkelsen, X. Xie, M. Kamal, D.J. Huebert, J. Cuff, B. Fry, A. Meissner, M. Wernig, K. Plath, R. Jaenisch, A. Wagscha, R. Feil, S.L. Schreiber, and E.S. Lander. A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell*, 125:315–326, 2006.

[84] D. Karolchik, R. Baertsch, M. Diekhans, T.S. Furey, A. Hinrichs, Y.T. Lu, K.M. Roskin, M. Schwartz, C.W. Sugnet, D.J. Thomas, R.J. Weber, D. Haussler, and W.J. Kent. The UCSC genome browser database. *Nucleic Acids Res*, 31:51–54, 2003.

[85] P.A. Fujita, B. Rhead, A.S. Zweig, A.S. Hinrichs, D. Karolchik, M.S. Cline, M. Goldman, G.P. Barber, H. Clawson, A. Coelho, M. Diekhans, T.R. Dreszer, B.M. Giardine, R.A. Harte, J. Hillman-Jackson, F. Hsu, V. Kirkup, R.M. Kuhn, K. Learned, C.H. Li, L.R. Meyer, A. Pohl, B.J. Raney, K.R. Rosenbloom, K.E. Smith, D. Haussler, and W.J. Kent. The UCSC genome browser database: update 2011. *Nucleic Acids Res.*, 39(Database issue):D876–82, 2011.

[86] M.E. Higgins, M. Claremont, J.E. Major, C. Sander, and A.E. Lash. Cancergenes: a gene selection resource for cancer genome projects. *Nucleic Acids Res.*, 35(Database issue):D721–6, 2007.

[87] S. Fan and X. Zhang. CpG island methylation pattern in different human tissues and its correlation with gene expression. *Biochemical and Biophysical Research Communications*, 383:421–5, 2009.

[88] `http://www.sanger.ac.uk/genetics/CGP`.

[89] J.L. Huret, P. Dessen, and A. Bernheim. An internet database on genetics in oncology. *Oncogene*, 22:1907, 2003.

[90] A. Hamosh, A.F. Scott, J.S. Amberger, C.A. Bocchini, and V.A. McKusick. Online mendelian inheritance in man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.*, 33(Database issue):D514–7, 2005.

[91] X. She, C.A. Rohl, J.C. Castle, A.V. Kulkarni, J.M. Johnson, and R. Chen. Definition, conservation and epigenetics of housekeeping and tissue-enriched genes. *BMC Genomics*, 10:269, 2009.

[92] L. Shen, Y. Kondo, Y. Guo, J. Zhang, L. Zhang, S. Ahmed, J. Shu, X. Chen, R.A. Waterland, and J.P. Issa. Genome-wide profiling of DNA methylation reveals a class of normally methylated CpG island promoters. *PLoS Genet.*, 3:2023–36, 2007.

[93] H. Zheng and H. Wu. Enhancement on the predictive power of the prediction model for human genomic DNA methylation. *International Conference on Bioinformatics and Computational Biology (BIOCOMP)*, 2011.

[94] Y. Zhang, H. Liu, J. Lv, X. Xiao, J. Zhu, X. Liu, J. Su, X. Li, Q. Wu, F. Wang, and Y. Cui. QDMR: a quantitative method for identification of differentially methylated regions by entropy. *Nucleic Acids Res.*, 39:e58, 2011.

[95] S. Schbath, B. Prum, and E. Turckheim. Exceptional motifs in different markov chain models for a statistical analysis of DNA sequences. *Journal of Computational Biology*, 2:417–437, 1995.

[96] J. Goñi, A. Pérez, D. Torrents, and M. Orozco. Determining promoter location based on DNA structure first-principles calculations. *Genome Biology*, 8:R263, 2007.

[97] N. Kaplan, I.K. Moore, Y. Fondufe-Mittendorf, A.J. Gossett, D. Tillo, Y. Field, E.M. LeProust, T.R. Hughes, J.D. Lieb, J. Widom, and E. Segal. The DNA-encoded nucleosome organization of a eukaryotic genome. *Nature Letter*, 458:362–366, 2009.

[98] S. Fan, F. Fang, X. Zhang, and M.Q. Zhang. Putative zinc finger protein binding sites are over-represented in the boundaries of methylation-resistant CpG islands in the human genome. *PLoS ONE*, 2:e1184, 2007.

[99] P.A. Futreal, L. Coin, M. Marshall, T. Down, T. Hubbard, R. Wooster, N. Rahman, and M.R. Stratton. A census of human cancer genes. *Nature Reviews Cancer*, 4:177–183, 2004.

[100] A. Barski, S. Cuddapah, K. Cui, T. Roh, D.E. Schones, Z. Wang, G. Wei, I. Chepelev, and Zhao K. High-resolution profiling of histone methylations in the human genome. *Cell*, 129:823–837, 2007.

[101] Z. Wang, C. Zang, J.A. Rosenfeld, D.E. Schones, A. Barski, S. Cuddapah, K. Cui, T. Roh, W. Peng, M.Q. Zhang, and K. Zhao. Combinatorial patterns of histone acetylations and methylations in the human genome. *Nature Genetics Letter*, 40:879–903, 2008.

[102] Y. Wu and A. Zhang. Feature selection for classifying high-dimensional numerical data. *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2:251–258, 2004.

[103] A. Agresti. A survey of exact inference for contingency tables. *Proceedings of the National Academy of Sciences USA*, 7:131–153, 1992.

[104] N. Turner. Chi-squared test. *Journal of Clinical Nursing*, 9:93, 2000.

[105] F. Yates. Contingency table involving small numbers and the χ^2 test. *Journal of the Royal Statistical Society*, 1:217–235, 1934.

[106] G. Marsaglia, W. Tsang, and J. Wang. Evaluating kolmogorov's distribution. *Journal of Statistical Software*, 8:1–4, 2003.

[107] J.V. Freeman and S.A. Julious. The analysis of categorical data. *Scope*, 16:18–21, 2007.

[108] J.V. Freeman and M.J. Campbell. The analysis of categorical data: Fisher's exact test. *Scope*, 16:18–21, 2007.

[109] K. Zhang, J.S. Siino, P.R. Jones, P.M. Yau, and E.M. Bradbury. A mass spectrometric western blot to evaluate the correlations between histone methylation and histone acetylation. *Proteomics*, 4:3765–3775, 2004.

[110] I.T. Jolliffe. Principal component analysis. *Springer-Verlag*, page 487, 1986.

[111] M.E. Wall, A. Rechtsteiner, and L.M. Rocha. *A Practical Approach to Microarray Data Analysis*, chapter Singular value decomposition and principal component analysis, pages 91–109. Kluwer: Norwell, MA, 2003.

[112] H.P. Kriegel, P. Kroger, E. Schubert, and A. Zimek. A general framework for increasing the robustness of PCA-based correlation clustering algorithms. *Scientific and Statistical Database Management*, 5069:418, 2008.

[113] A.J. Butte, V.J. Dzau, and S.B. Glueck. Further defining housekeeping, or "maintenance," genes focus on "a compendium of gene expression in normal human tissues". *Physiol Genomics*, 7:95–6, 2001.

[114] R.M. Kuhn, D. Karolchik, A.S. Zweig, T. Wang, K.E. Smith, K.R. Rosenbloom, B. Rhead, B.J. Raney, A. Pohl, M. Pheasant, L. Meyer, F. Hsu, A.S. Hinrichs, R.A. Harte, B. Giardine, P. Fujita, M. Diekhans, T. Dreszer, H. Clawson, G.P. Barber, D. Haussler, and W.J. Kent. The UCSC genome browser database: update 2009. *Nucleic Acids Res*, 37:D755–61, 2009.

[115] C.W. Hsu, C.C. Chang, and C.J. Lin. A practical guide to support vector classification. *Technical Report*, 2003.

[116] B. Li and M. Meng. Tumor recognition in wireless capsule endoscopy images using textural features and SVM-based feature selection. *IEEE Trans Inf Technol Biomed.*, 99:1, 2012.

[117] C.M. Koch, R.M. Andrews, P. Flicek, S.C. Dillon, U. Karaoz, G.K. Clelland, S. Wilcox, D.M. Beare, J.C. Fowler, P. Couttet, K.D. James, G.C. Lefebvre, A.W. Bruce, O.M. Dovey, P.D. Ellis, P. Dhami, C.F. Langford, Z. Weng, E. Birney, N.P. Carter, D. Vetrie, and I. Dunham. The landscape of histone modifications across 1% of the human genome in five human cell lines. *Genome Res*, 17:691–707, 2007.

[118] M. Esteller. Aberrant DNA methylation as a cancer-inducing mechanism. *Annu. Rev. Pharmacol.Toxicol*, 45:629–656, 2005.

[119] I. Ali and H. Seker. A comparative study for characterisation and prediction of tissue-specific DNA methylation of CpG islands in chromosomes 6, 20 and 22. *Conf Proc IEEE Eng Med Biol Soc.*, pages 1832–5, 2010.

[120] R.A. Irizarry, C. Ladd-Acosta, B. Wen, Z. Wu, C. Montano, P. Onyango, H. Cui, K. Gabo, M. Rongione, M. Webster, H. Ji, J.B. Potash, S. Sabunciyan, and A.P. Feinberg. The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores. *Nature Reviews Cancer*, 41:178–186, 2009.

[121] P. Nawapen, S. Junpen, H. Dion, D. Michael, C. Bernie, and T. Mongkol. Different DNA methylation patterns detected by the Amplified Methylation Polymorphism

Polymerase Chain Reaction (AMP PCR) technique among various cell types of bulls. *Acta Veterinaria Scandinavica*, 52:18, 2010.

[122] M. Szyf. Therapeutic implications of DNA methylation. *Future Oncology*, 1:125–35, 2005.

[123] P.A. Abreu, G. Dellamora-Ortiz, L.R. Leao-Ferreira, M. Gouveia, E. Braggio, I. Zalcberg, D.O. Santos, S. Bourguinhon, L.M. Cabral, C.R. Rodrigues, and H.C. Castro. DNA methylation: a promising target for the twenty-first century. *Expert Opin Ther Targets*, 12:1035–47, 2008.

VITA

Hao Zheng was born in 1983 in Shanghai, China. He received his B.Eng. degree in Shanghai Jiao Tong University, Shanghai, China, in 2006. He received his M.S. degree in the School of Electrical and Computer Engineering from Georgia Institute of Technology in 2007. Since the spring of 2008, he has been continuing to pursue the Ph.D. degree in the School of Electrical and Computer Engineering, Georgia Institute of Technology. His research interests include machine learning, signal processing and informatics with applications in biological and medical sciences.

Publications

- Hao Zheng, S.W. Jiang, and H. Wu, A Review on the Techniques for Characterizing and Predicting Human Genomic DNA Methylation, accepted, Current Bioinformatics, 2012.
- Hao Zheng, S.W. Jiang, and H. Wu, Enhancement on the Predictive Power of the Prediction Model for Human Genomic DNA Methylation, International Conference on Bioinformatics and Computational Biology (BIOCOMP), 2011.
- Hao Zheng and H. Wu, Gene-centric Association Analysis for the Correlation between the Guanine-Cytosine Content Levels and Temperature Range Conditions of Prokaryotic Species, BMC Bioinformatics, 11:S7, 2010
- Hao Zheng and H. Wu, Short Prokaryotic DNA Fragment Binning Using a Hierarchical Classifier Based on Linear Discriminant Analysis and Principal Component Analysis, Journal of Bioinformatics and Computational Biology, 8(6):995-1011, 2010.
- Hao Zheng and H. Wu, Analysis on the Correlation Relationships between the Temperature Range Condition and the Genic GC Content Levels of Prokaryotes, IEEE

International Conference on Bioinformatics and Bioengineering (BIBE), 2010.

- Hao Zheng and H. Wu, A Novel LDA and PCA-based Hierarchical Scheme for Metagenome Fragment Binning, IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), 2009.